An Introduction to

ECONOMETRICS

Jaydeb Sarkhel
Santosh Kumar Dutta

Written according to the new C.B.C.S. syllabus of B.A./B.Sc. Honours Course in Economics of different universities in India.

AN INTRODUCTION TO
ECONOMETRICS

B.A./B.Sc. Economics (Honours)

Jaydeb Sarkhel

Retired Professor, Department of Commerce,
Burdwan University;
Author of 'Microeconomic Theory', 'Macroeconomic Theory', etc.

Dr. Santosh Kumar Dutta

Associate Professor, Department of Economics,
Bidhannagar College;
Joint Author of 'An Insight into Statistics',
'Microeconomics and Statistics' etc.

Revised and Enlarged
Second Edition
2020

BOOK SYNDICATE (P) LTD.
www.booksyndicate.in

CBCS (UG) Syllabus on Econometrics for different Universities in West Bengal

Paper 1.3 CALCUTTA UNIVERSITY

Economics Core Course X, Core T_{10}
–Introductory Econometrics (Sem-IV), F.M : 100
(Th : 60 + Tutorial : 20 + Internal assessment : 10 + Attendence : 10)

1. Nature and Scope of Econometrics **3 Lecture hours**
1.1 What is Econometrics ?
1.2 Distinction between Economic model and Econometric model
1.3 Concept of Stochastic relation
1.4 Role of random disturbance in econometric model
2. Classical Linear Regression Model (simple linear regression and multiple linear regression) : Part 1 **15 Lecture hours**
2.1 The classical assumptions
2.2 Concepts of population regression function and sample regression function
2.3 Estimation of model by the method of ordinary least squares
3. Classical linear regression model (simple linear regression and multiple linear regression) : Part 2 **15 Lecture hours**
3.1 Properties of Least Squares Estimators (BLUE)-Gauss-Markov theorem
3.2 Qualitative (dummy) independent variables (only interpretation of the model)
3.3 Forecasting (only for two variable model) : Expost forecast and Exante forecast
4. Statistical inference in Linear regression model **20 Lecture hours**
4.1 Sampling distribution of regression estimates : Standard normal, Chi-square, t, F
4.2 Confidence intervals
4.3 Concepts of Type I and Type II errors
4.4 Testing of hypothesis about β and σ^2 given and with unknown σ^2 (Standard normal and 't' statistics)
4.5 Testing hypothesis involving several parameters : the F test
4.6 Goodness of fit (in terms of R^2, adjusted R^2 and F statistic)
5. Violations of Classical Assumptions **10 Lecture hours**
5.1 Multicollinearity–Consequences, Detection and Remedies
5.2 Heteroscedasticity–Consequences, Detection and Remedies
5.3 Autocorrelation–Consequences, Detection and Remedies
6. Specification Analysis **10 Lecture hours**
6.1 Omission of a relevant variable
6.2 Inclusion of an irrelevant variable
6.3 Tests of specification errors
6.4 Testing for linearity and normality assumptions.

CONTENTS

Chapter 3 **Multiple Linear Regression Model** **105–184**

Chapter 5 Specification Analysis

1

Definition, Scope and Goals of Econometrics

1.1. Definition and Scope of Econometrics

Literally speaking, the word 'econometrics' means "measurement in economics". Econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories. It is a special type of economic analysis in which the general economic theory formulated in mathematical terms is combined with empirical measurement of economic phenomena. We start from general economic theory, that is, from the relationships of economic variables as suggested by economic theory and express them in mathematical terms. This is called building of an economic model. Next we use statistical methods in order to obtain numerical estimates of the coefficients of the economic relationships. These statistical methods are called ***econometric methods.***

Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations :

"Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results"[1].

"Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference."[2]

"Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena."[3]

"Econometrics is concerned with the empirical determination of economic laws."[4]

"The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him."[5]

"Econometricians ... are of positive help in trying to dispel the poor public image of economics (quantitative or otherwise) as a subject in which empty boxes are opened

---

[1] Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics*, The University of Chicago Press, Chicago, 1968, p. 74.

[2] P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committe for Econometrics," *Econometrica*, vol. 22, no. 2, April 1954, pp. 141-146.

[3] Arthur S. Goldberger, *Econometric Theory*, John Wiley & Sons, New York, 1964, p. 1.

[4] H. Theil, *Principles of Econometrics*, John Wiley & Sons, New York, 1971, P. 1.

[5] E. Malinvaud, *Statistical Methods of Econometrics*, Rand Mc Nally, Chicago, 1966, P. 514.

by assuming the existence of counterparts to several concepts which any two economists will interpret in 11 ways."[6]

"The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier."[7]

All these definitions suggest that econometrics is an amalgam of economic theory, mathematical economics, economic statistics and mathematical statistics.

Economic theory postulates an exact relationship between economic variables but actually an economic relationship always contains a random element. Economic theory ignores it but econometrics does not, because the econometric methods can deal with these random components. For example, in the Keynesian macroeconomic theory we find an exact relationship between consumption expenditure (C) and income (Y). Keynesian consumption function is given by $C = a + bY$ where $a > 0$ is called the autonomous part of consumption expenditure and $b = \frac{dC}{dY}$ is called the marginal propensity to consume (MPC) [$0 < b < 1$ by assumption]. This is an exact relationship because C is completely determined by Y. So, in this model the effects of other variables like price, wealth, income distribution etc., are ignored. But in econometrics the influence of other factors is considered by introducing a random variable in the model and that random variable is generally denoted by 'u' called error term. So, the consumption function considered in econometrics is $C = a + bY + u$. Now econometric methods estimate the parameters 'a' and 'b' and while estimating 'a' and 'b' the choice of the econometric method depends on the behaviour of the distribution of the random variable 'u'.

There are three main sources of the error term 'u' in the functional relation. These are :

(i) unpredictable element of randomness in human response,

(ii) effect of a large number of variables that have been omitted from the functional relation

and (iii) measurement error. (For details see Section 2.2.1)

1.2. Relationship between Econometrics and Economic Theory

Economic theory makes statements or hypotheses that are mostly qualitative in nature.

Econometrics presupposes the existence of a body of economic theory. Economic theory should come first because it states the hypothesis about economic behaviour which should be tested with the econometric methods.

For example, we consider the consumption income relationship of the form

$$C = a + bY + u.$$

Economic theory suggests that consumption is a function of income and with the information of economic theory we know that $MPC = \frac{dC}{dY} = b$ lies between 0 and 1 i.e., $0 < b < 1$.

---

[6] Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics*, Edward Elgar Publishing, Hants, England, 1990, P. 54

[7] T. Haavelmo, "The Probability Approach in Econometrics", Supplement to *Econometrica*, vol. 12, 1944, Preface P. iii.

The proposition suggested by economic theory is to be tested now, applying econometric methods. If we find that the theory is consistent with the empirical results we accept the theory but if we find that it is not consistent with the empirical results, then we have either to reject the theory or to modify the theory. If we like to modify the theory then we should not reject the theory, rather we should incorporate some other variables and parameters to make the theory more meaningful (and close to reality).

For example, the simple consumption income relation, $C = a + bY + u$ can be modified in the form

$C = a + bY + cP + dW + u$ where two new variables P (price level) and W (wealth) have been taken into account in the functional relation. The signs of the parameters (a, b, c, $d > 0$) and the corresponding response coefficients can also be tested empirically.

1.2.1. Difference between Economic Model and Econometric Model

A model is a simplified representation of a real world process. In practice, in any economic model (say consumption function or demand function), we can include all the relevant variables that we think are relevant for our purpose and dump the rest of the variables in a basket called "disturbance". This brings us to the distinction between an economic model and an econometric model.

An economic model is a set of assumptions that approximately describe the behaviour of an economy. An econometric model, on the other hand, consists of the following :

(i) A set of behavioural equations derived from the economic model.

(ii) A statement of whether there are errors of observation in the observed variables.

(iii) A specification of the probability distribution of the "disturbances" (and errors of measurement).

For example, we may consider a simple demand model of economics. Then econometric model will usually consist of :

(a) The behavioural equation : $q = \alpha + \beta p + u$ where q = quantity demanded, p = price, α and β are two parameters and u = random disturbance term.

(b) A specification of probability distribution of u, where values of u are independently and normally distributed with mean $E(u) = 0$ and variance $(u) = \sigma_u^2$. With these specifications we can test empirically the law of demand or the hypothesis that $\beta < 0$.

We may also use the estimated demand function for prediction and policy purposes.

1.3. Econometrics and Mathematical Economics

Mathematical economics states economic theory in terms of mathematical symbols. There is no essential difference between mathematical economics and economic theory. Both state the same relationships, but while economic theory uses verbal exposition, mathematical economics employs mathematical symbolism. Both express the various economic relationships in an exact form. Neither economic theory nor mathematical economics allows for random elements which might affect the relationship and make it *stochastic*. Furthermore, they do not provide numerical values for the coefficients of the relationships. Relations in economic theory or in mathematical economics are of *non-stochastic* form. It is in this regard that econometrics differs from mathematical economics.

Although econometrics presupposes the expression of economic relationships in mathematical form, like mathematical economics it does not assume that economic relationships are exact. On the contrary, econometrics assumes that relationships are not exact. Econometric methods are designed so take into account random disturbances which create deviations from the exact behavioural patterns suggested by economic theory and mathematical economics. Furthermore, econometric methods provide numerical values of the coefficients of economic phenomena. Thus, by combining mathematical formulations of theory with empirical data, econometrics enables us to pass from the abstract theoretical scheme to numerical results in concrete cases.

1.4 Econometrics and Statistics

Econometrics differs both from mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or charts them, and then attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. Thus economic statistics is mainly a descriptive aspect of economic theory. It does not provide explanations of the development of the various variables and it does not provide measurement of the parameters of economic relationships.

Economic statistics differs from mathematical (or inductive) statistics. Mathematical statistics is based upon the theory of probability and deals with the methods of measurement which are developed on the basis of controlled or carefully planned experiments. These statistical methods cannot be applied to economic relationships because such experiments cannot be designed, except in a very few cases, e.g. agricultural experiments or industrial experimentation, for economic phenomena.

Econometrics uses statistical methods after adapting them to the problems of economic life. These adapted statistical methods are called econometric methods. In particular, econometric methods are so adjusted that they become appropriate for the measurement of economic relationships which are stochastic, that is, they include random elements. The adjustment consists primarily in specifying the stochastic (random) elements that are supposed to operate in the real world and enter into the determination of the observed data, so that the latter can be interpreted as a (random) sample to which the methods of statistics can be applied.

1.5. Goals of Econometrics

Econometrics helps us to achieve the following three main goals:

(i) **Analysis**: This means testing of economic theory. There are alternative theories to explain the functioning of the economic system. Econometrics examines the explanatory power of the system.

(ii) **Policy making**: The numerical estimates of the coefficients of the economic relationships help the policy-maker to define the appropriate policies. For example, the numerical estimate of price elasticities of demand for a product will help the policy maker to know how much additional revenue is expected to be obtained if sales tax is imposed on that commodity. Alternatively, numerical estimates of price elasticities of exports and imports will help us to know how far the devaluation as a policy will be effective in solving the balance of payments deficit problem.

(iii) **Forecasting**: The numerical estimates of the coefficients are used in order to forecast the future value of the economic variables. Without forecasting the planner cannot adopt appropriate policies. Of course, these goals are not mutually exclusive.

Successful econometric applications should really include some combination of all these aims.

1.6 Division of Econometrics

Econometrics may be divided into two branches: theoretical econometrics and applied econometrics.

Theoretical econometrics includes the development of appropriate methods for the measurement of economic relationships. Econometric techniques are basically statistical techniques which have been adapted to the particular characteristics of economic relationships.

Econometric methods may be classified into two groups: (i) Single-equation techniques which are methods that are applied to one relationship at a time, and (ii) Simultaneous equation techniques, which are methods applied to all the relationships of a model simultaneously.

Applied econometrics includes the applications of econometric methods to specific branches of economic theory. It examines the problems encountered and the findings of applied research in the fields of demand, supply, production, investment, consumption and other sectors of economic theory. Applied econometrics involves the application of the tools of theoretical econometrics for the analysis of economic phenomena and forecasting economic behaviour.

1.7. Methodology/Stages of Econometric Research

Applied econometric research is concerned with the measurement of the parameters of economic relationships and with the prediction of the values of economic variables.

The relationships of economic theory which can be measured with one or another econometric techniques are causal, that is, they are relationships in which some variables are postulated as causes of the variation of other variables. In this sense definitional equations do not require any measurement. For example, the equation $Y = C + I$ is the mathematical expression of the definition of national income (in a closed economy with no government activity) of economic theory. It does not explain the determinants of the level of income or the causes of its variations.

There are four stages in any econometric research.

Stage A: Specification of the model.

It means expressing the relationships between the variables in mathematical form. This stage is also called formulation of the maintained hypothesis. It involves the determination of

(i) dependent and the explanatory variables to be included in the model

(ii) the theoretical expectations about the sign, size of the parameters of the function

(iii) the mathematical form of the model

For example, consider a production function of the following type $Y = f(K, L)$ where K and L are the two factors of production.

K = Capital, L = Labour and Y is the level of output.] This function can also be written in the Cobb-Douglas form i.e. $Y = K^{\alpha} L^{\beta}$ or $\log Y = \alpha \log K + \beta \log L$. This is the mathematical form of log linear function. Here some theoretical restrictions must be imposed: $0 < \alpha, \beta < 1$,

$\alpha + \beta > 1$ if there are increasing returns to scale

$\alpha + \beta < 1$ if there are decreasing returns to scale

$\alpha + \beta = 1$ if there are constant returns to scale α = Elasticity of output with respect to capital

β = Elasticity of output with respect to labour

Stage B Estimation of the model

[illegible]

[illegible] the variables of the function [illegible] among the explanatory variables [illegible] $\beta_1 P + \beta_2 Y + \beta_3 W + u$ where [illegible] dependent variables and explanatory variables [illegible] we have [illegible] the problem of multicollinearity.)

[illegible] the appropriate econometric technique for the estimation [illegible] function and critical examination of the assumptions of the chosen technique [illegible] their economic implications for the estimates of the coefficients.

Stage C Evaluation of estimates

After the estimation of the model the econometrician must proceed with the evaluation of the results of the calculations, that is with the determination of the reliability of these results. The evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically satisfactory.

For this purpose we may use various criteria which may be classified into three main groups.

(i) **Economic criteria** These are determined by the principles of economic theory and refer to the sign and the size of the parameters of economic relationships. For example, the Keynesian liquidity preference function may be expressed in the mathematical form

$$M = \beta_0 + \beta_1 Y + \beta_2 r + u$$

where M = demand for money (dependent variable), Y = income, r = rate of interest, u = error term, β_0, β_1, β_2 are the parameters whose values and signs are to be determined on the basis of observed data. On the basis of the existing theory, the signs of the parameters would be: $\beta_0 > 0$, $\beta_1 > 0$, $\beta_2 < 0$

(ii) **Statistical criteria (First order tests)** These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. The most widely used statistical criteria are the correlation coefficient and the standard error of the estimates.

(iii) **Econometric criteria (Second order tests)** These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as second order tests (as tests of the statistical tests); in other words they determine the reliability of the statistical criteria, and in particular the standard errors of the parameter estimates. [illegible] help us establish whether or [illegible]

Stage D Evaluation of the forecasting power of the estimated model

[illegible]

[illegible] and test the forecasting power of the model.

1.8 Desirable Properties of an Econometric Model

The goodness of an econometric model [illegible] following desirable properties:

(i) **Theoretical plausibility** The model must be compatible with the postulates of economic theory. It must describe the [illegible] to which it relates.

(ii) **Explanatory ability** The model should be able to explain [illegible] the actual world. It must be consistent with the observed behaviour of the economic variables whose relationship it determines.

(iii) **Accuracy of the estimates of the parameters** The estimates of the coefficients should be accurate in the sense that they should approximate as best as possible the true parameters of the structural model.

(iv) **Forecasting ability** The model should produce satisfactory predictions of future values of the dependent variables.

(v) **Simplicity** The model should represent the economic relationships with maximum simplicity.

1.9 Nature and Sources of Data for Economic Analysis

The success of any econometric analysis depends on the availability of the appropriate data. Three types of data are generally available for empirical analysis: *time series data*, *cross section data* and *pooled data (panel data)*.

Time Series Data

A time series is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as *daily* (e.g., stock prices, weather reports), *weekly* (e.g., money supply figures), *monthly* (e.g., unemployment rate, Consumer Price Index (CPI)), *quarterly* (e.g., GDP), *annually* (e.g., government budget), *quinquennially*, that is, every 5 years (e.g., the census of manufactures) or *decennially*, that is every 10 years (e.g., the census of population).

Cross-Section Data

Cross-section data are data on one or more variables collected at the same point of time, such as the census of population conducted by the Government of India every 10 years, the Survey of household consumer expenditure in India conducted by National Sample Survey Organization (NSSO), the opinion polls by the Times of India, NDTV, CNN-IBN and many other organizations. An individual researcher or a group may also collect cross-section data directly from the field of enquiry.

Conventionally, the letter Y denotes the dependent variable and X's ($X_1, X_2, \ldots, X_k$) denote the explanatory (independent) variables, X_k being the kth explanatory variable. The subscript i or t denote ith or the tth observation or value. X_{ki} (or X_{kt}) will denote the ith (or tth) observation on variable X_k. Here N (or T) will denote the total number of observations or values in the population and n (or t) will denote the total number of observations in a sample. Normally the subscript i will be used for cross-section data (i.e. data collected at one point of time) and the subscript t will be used for time series data (i.e., data collected on different periods of time). For instance, consider the Keynesian consumption function of the form $C = a + bY$ where C = consumption expenditure, Y = income and a and b are two constants, a = autonomous part of consumption expenditure, b = marginal propensity to consume. According to the existing theory $a > 0$, $0 < b < 1$. If we like to test this relation with the help of time series data, then we will write the regression equation in the form $C_t = a + bY_t + u_t$ (where $t = 1, 2, \ldots, T$ (say)) where u is the random disturbance term. On the other hand we can write the regression equation in the form $C_i = a + bY_i + u_i$ ($i = 1, 2, \ldots, N$ say), when we verify the relation with the help of cross-section data.

Pooled Data

Pooled or combined data are elements of both time series and cross section data. Generally speaking, pooled data is a combination of data (i.e. sales, advertisement, earnings etc.) of say 20 firms over a given period of time say a year or two. These combined data of 20 firms in 2 years making 40 observations make it a pooled data, that is, pooling 20 firms' data in 2 years together. So, it is a combination of cross section data and time series data.

Panel, Longitudinal, or Micropanel Data

This is a special type of pooled data in which the same cross-sectional unit (say a family or a firm) is surveyed over time.

For example, the Nigerian population commission surveys each house every 10 years to determine the changes that may have occurred within these years. By surveying or interviewing the same households or firms to find out their population or financial conditions periodically (10 years interval), panel data can help to provide useful information on the changes that may have occurred in these households. It is more detailed than just the pooled data in a short period of time.

Thus, there is a basic difference between pooled data and panel data. It should be noted that pooled time-series, cross-section data are data with relatively few cross-sections (say few firms under study), where variables are held in cross-section specific individual series (i.e. sales, advertisement, earnings, etc.), while panel data correspond to data with large number of cross-sections, with variables held in single series in stacked form.

The Sources of Data

The data used in empirical analysis may be collected by a government agency (e.g. the Central Statistical Organization), an international agency (e.g., the International Monetary Fund (IMF) or the World Bank), a private organization (e.g. the Centre for Monitoring Indian Economy) or an individual. There exist a lot of agencies collecting data for one purpose or another. Now a days the Internet has revolutionized data gathering. Most of the data can be downloaded from different websites either free of cost or with minimum cost.

The Accuracy of Data

Although plenty of data are available for economic research, the quality of data is often not that good.

There are several reasons for that:

(i) Most of the social science data are non experimental in nature. Therefore there is the possibility of observational errors.

(ii) Even in experimentally collected data, errors of measurement arise from approximations and rounding offs.

(iii) In questionnaire type of surveys, the problem of non-response may lead to bias in results.

(iv) The sampling methods used in obtaining data may vary so widely that it is often difficult to compare the results obtained from the various samples.

(v) Economic data are generally available at a highly aggregate level. Such highly aggregated data may not be helpful for individualistic study.

Because of all of these and many other problems, the researchers should always keep in mind that the results of research are only as good as the quality of the data. Therefore, if in given situations researchers find that the results of the research are "unsatisfactory" the cause may not be that they used the wrong model but due to the poor quality of data.

1.10. A Note on the Measurement Scales of Variables

The variables that we generally use can be measured in four types of scales: *ratio scale, interval scale, ordinal scale* and *nominal scale*. We can briefly describe them as follows:

Ratio Scale: For a variable X, taking two values say X_1 and X_2, the ratio X_1/X_2 and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a natural ordering (ascending or descending) of the values along the scale (say $X_2 \geq X_1$ or $X_2 \leq X_1$). Most economic variables belong to this category. Personal income, measured in rupees is a ratio variable, someone earning ₹ 50,000 is making twice as much as another person earning ₹ 25000.

Interval scale : The interval scale satisfies the last two properties stated in ratio scale, but not the first.

For example, the distance between two time periods say (2018–2001) is meaningful but not the ratio of two time perids $\left(\frac{2018}{2001}\right)$

Ordinal Scale: A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering). Examples are grading systems (A, B, C, grades) or income class (upper, middle, lower). For these variables the ordering exists. But the distances between the categores cannot be quantified.

Nominal Scale: Variables in this category have none of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories.

EXERCISE

1. What is Econometrics and what are its components ? Describe the function of each component. Give examples in support of your answer.
2. Name and describe three relationships studied in Economic theory which can be estimated as subject-matter of Econometrics. What are the parameters of these relationships ?
3. How would you define Econometrics ? How does it differ from Mathematical Economics and Statistics ? Describe the main steps involved in any econometric research by taking an example from economic theory.
4. Considering the following relations, how would you explain that economic theory postulates exact relationships between economic variables. How can these be transformed into econometric relations ?

   (i) Demand function : $D = \alpha + \beta P + \gamma I$ where D = quantity demanded, P = price and I = income.

   (ii) Supply function : $S = \alpha + \beta P$ where S = quantity supplied, P = price.

   (iii) Consumption function : $C = \alpha + \beta Y_d$ where C = consumption expenditure and Y_d = disposable income.

   (iv) Cost function : $C = \alpha + \beta X$ where C = total cost and X = total output.

   (v) Production function : $Y = AL^{\alpha} K^{\beta}$ where Y = level of output, L = labour input, K = capital input, A = constant technical parameter; α, β are the two elasticity coefficients.

   (a) What is the economic meaning of the coefficients involved in all the above relations ?

   (b) What would you expect about the sign and size of the coefficients involved in each of the above relationships ?
5. Enumerate the relation between econometrics and economic theory.
6. What is Econometrics ? What are the different goals of econometrics ?
7. Distinguish between theoretical econometrics and applied econometrics.
8. Explain briefly the different stages of any econometric research.
9. What is an econometric model ? Illustrate any one of such models.
10. What are the desirable properties of an econometric model ?
11. What are the different types of data available for empirical analysis ?
12. Distinguish between time series data and cross-section data.
13. What are pooled data ? What are panel data ? Distinguish between pooled data and panel data.
14. What are the different sources of data used in empirical analysis.
15. What do you mean by accuracy of data ? What are the different reasons for distortion of accuracy of data collected and published by different organizations.
16. Give a brief outline on measurement scales of variables.

2

The Simple Linear Regression Model

2.1. Introduction

Most of economics is concerned with relations among variables. These relations when phrased in mathematical terms can predict the effect of one variable on another. For example, assuming that income, prices of other commodities and all other determinants of demand are constants, we can express the quantity demanded (q) of any commodity as a function of the price (p) of that commodity only. This may be put in the form $q = f(p)$. Similarly we are familiar with other functions (with different assumptions) such as consumption function ($C = f(Y)$), supply function ($S = f(p)$), cost function ($C = f(q)$), production function ($Q = f(x_1, x_2)$) where x_1 and x_2 are amounts of different inputs, etc.

These functional relationships define the dependence of the dependent variable upon the independent variable(s) in the specific form. The functional relation may be linear, quadratic, logarithmic, exponential or hyperbolic.

A relation between two variables X and Y expressed as $Y = f(X)$ is said to be deterministic or non-stochastic (non-random) if for each value of the independent variable (X) there is one and only one corresponding value of the dependent variable (Y). On the other hand, a relation between X and Y is said to be stochastic if for a particular value of X there is a whole probability distribution of values of Y. In such a case for any given value of X, the dependent variable Y assumes some specific value only with some probability.

For example, a linear demand function (in deterministic form) can be written as $q = f(p) = \alpha + \beta p$ ($\alpha > 0$, $\beta < 0$) and in particular $q = 100 - 5p$. When $p = 10$, $q = 50$, when $p = 15$, $q = 25$ etc.

But such an exact and deterministic relation between p and q is never true in the real world.

The deterministic behaviour of the above relationship breaks down when the *ceteris paribus* (other things remaining the same) condition is relaxed.

We therefore rewrite the demand equation as $q = \alpha + \beta p + u$ or in particular $q = 100 - 5p + u$ where u is commonly known as *random disturbance* since it disturbs an otherwise deterministic relation.

2.1.1. Concepts of Population Regression Function and Sample Regression Function

Sampling denotes the selection of a part of the aggregate statistical material with a view to obtaining information about the whole. This aggregate or totality of statistical information on a particular character of all the members covered by an investigation is called population or universe. When the population size is very large, it may not be possible to take a complete enumeration of the population. Then we select a small part of the population which is called sample and examining this small part we infer about the nature of the whole population. The basic objective of sampling is to make inference about the population by examining a small part of it.

In regression we may be interested to find out the relation between two or more variables simultaneously. In the case of simple (linear) regression model we assume only one explanatory variable but in the case of multiple regression model we assume more than one explanatory variables. The first case is known as the bivariate analysis while the second case is known as the multivariate analysis. Initially we will concentrate on bivariate analysis (study the relation between two variables X and Y only, where Y dependent variable, X independent explanatory variable).

We know that regression analysis is largely concerned with estimating and/or predicting the population parameter say mean value of the dependent variable (Y) on the basis of the known or fixed values of the explanatory variable(s). To understand this, we consider a total population of 60 families in a hypothetical community and their monthly income (X) and monthly consumption expenditure (Y), both in rupees. These 60 families are divided into 10 income groups and the monthly expenditures of each family in the various groups are shown in the following table (Table 2.1).

Table 2.1 Joint distribution of monthly income (X in ₹) and monthly consumption expenditure (Y in ₹) of 60 families in a hypothetical community

X → Y ↓	8000	10000	12000	14000	16000	18000	20000	22000	24000	26000
	5500	6500	7900	8000	10200	11000	12000	13500	13700	15000
	6000	7000	8400	9300	10700	11500	13600	14700	14500	15200
	6500	7400	9000	9500	11000	12000	14000	14000	15500	17500
	7000	8000	9400	10300	11600	13000	14400	15200	16500	17800
	7500	8500	9800	10800	11900	13500	14500	15700	17500	18000
	·	8800		11300	12500	14000		16000	18900	18500
		—		11500			—	16200		19000
Total	32500	46200	44500	70700	67900	75000	68500	104300	96600	121000
Conditional means of Y E(Y/X)	6500	7700	8900	10100	11300	12500	13700	14900	[illegible]	17400

Here we have 10 fixed values of X and the corresponding Y values against each of the X values and hence we have 10 subpopulations of Y. From Table 2.1, we see that there is considerable variation in monthly consumption expenditure in each income group but the general picture is that despite the variability of monthly consumption expenditure within each income bracket, on an average, monthly consumption expenditure increases as income increases. To understand it clearly we have given the mean, or average montly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the monthly income level of ₹8000, the mean consumption expenditure is ₹6500 and so on. In total we have 10 mean values for 10 sub-populations of Y and these mean values are called *conditional expected values* as they depend upon the given values of the (conditioning) variable X.

Symbolically, we denote them as $E(Y \mid X)$ which simply means the expected value of Y given the value of X. It should be noted that these expected values are called conditional expected values. In order to calculate the conditional expected values $E(Y \mid X)$ we have to construct conditional probability distribution of Y, $P(Y \mid X)$ as shown in Table 2.2.

Table 2.2 Conditional Probabilities $P(Y \mid X_i)$ for the data of Table 2.1

$X \rightarrow$ $P(Y \mid X_i) \downarrow$	8000	10000	12000	14000	16000	18000	20000	22000	24000	26000
Conditional probabilities $P(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	—	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
			—	$\frac{1}{7}$				$\frac{1}{7}$		$\frac{1}{7}$
Conditional means of Y	6500	7700	8900	10100	11300	12500	13700	14900	16100	17300

For the income group of ₹8000 the expected monthly expenditure is obtained by

$$₹5500 \times \frac{1}{5} + ₹6000 \times \frac{1}{5} + ₹6500 \times \frac{1}{5} + ₹7000 \times \frac{1}{5} + ₹7500 \times \frac{1}{5} = ₹6500$$

The expected monthly expenditures for other income groups are also obtained in this way.

It is important to distinguish these conditional expected values from the *unconditional expected value* of monthly consumption expenditure, $E(Y)$. If we add the monthly consumption expenditures for all the 60 families in the population and divide this number by 60, we get the value ₹12120 (₹727200/60) which is the unconditional mean or expected value of Y, $E(Y)$.

Thus the expected monthly consumption expenditure of a family would be ₹12120 (the unconditional mean). But if we like to know the expected value of monthly consumption expenditure of a family whose monthly income is say ₹12000, then we get a value of ₹8900 (the conditional mean).

Graphically, if we join these conditional mean values, we obtain the **population regression line (PRL)** or **population regression curve** or simply it is the *regression of Y on X*.

The population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More specifically, It is the curve which joins the means of the sub-populations of Y corresponding to the given values of the regressor X. This is shown in figure 2.1.

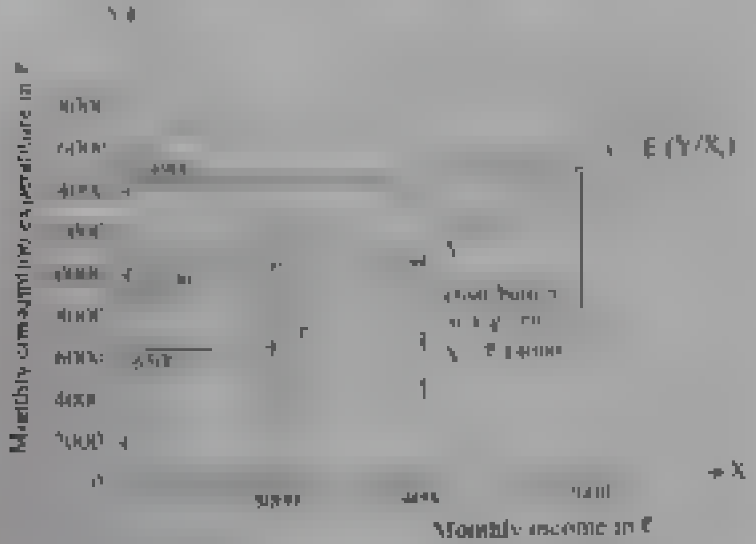


Fig. 2.1. Population regression line (data of Table 2.1)

This figure shows that for each X (i.e., income level) there is a population of Y values (monthly consumption expenditure) that are spread around the (conditional) mean of those Y values. For simplicity we have assumed that these Y values are distributed symmetrically around their respective (conditional) mean values and the regression line (or curve) passes through these (conditional) mean values.

2.1.2. Population Regression Function (PRF)

From the above explanation and Figure 2.1 it is clear that each conditional mean $E(Y/X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically $E(Y/X_i) = f(X_i)$ where $f(X_i)$ denotes some function of the explanatory variable X. $f(X_i)$ is a linear function in X_i. This function is also known as the **conditional expectation function (CEF) or Population regression function (PRF) or population regression (PR)**. It states merely that the expected value of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X. However the functional form of the PRF is an empirical question. For example, to verify the consumption income relation we generally assume a linear relation. We may assume that the PRF $E(Y/X_i)$ is a linear function of X_i, say, of the type

$E(Y/X_i) = \alpha + \beta X_i$ where α and β are unknown but fixed parameters known as the regression coefficients. α and β are also known as **intercept** and **slope** coefficients respectively. In regression analysis our interest is in estimating the PRF and the unknown values of α and β on the basis of observations on Y and X.

From our earlier example stated in Table 2.1 we see that, given the income level of X_i (say), an individual family's consumption expenditure is clustered around the average consumption of all families at X_i, i.e., around its conditional expectation. Therefore, we can express the deviation of an individual Y_i around its expected value as follows:

$$u_i = Y_i - E(Y/X_i) \text{ or } Y_i = E(Y/X_i) + u_i \text{ or } Y_i = \alpha + \beta X_i + u_i$$

where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is known as the **Stochastic disturbance** term or **Stochastic error term.**

2.1.3 The Sample Regression Function (SRF)

In the preceding [illegible] population [illegible] the [illegible] shown in Table [illegible] Table [illegible] the same [illegible] population [illegible]

Table 2.3

A random sample from the population of Table 2.1

Y	X
7000	8000
6500	10000
9000	12000
9500	14000
10000	16000
11500	18000
12000	20000
14000	22000
15500	24000
15000	26000

Table 2.4

A random sample from the population of Table 2.1

Y	X
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	22000
[illegible]	24000
[illegible]	26000

Now from the sample of Table 2.3 we can predict or forecast the average monthly consumption expenditure Y in the population as a whole corresponding to the chosen X's.

However, we may not be able to estimate the PRF accurately because of sampling fluctuations. To see this we have drawn another random sample from the same population (Table 2.1) and shown in Table 2.4.

Now plotting the data of Tables 2.3 and 2.4 we obtain the scatter diagram shown in Figure 2.2.

In the scatter diagram two sample regression lines are drawn so as to fit the scatters reasonably well. SRF_1 is based on the first sample and SRF_2 is based on the second sample. The regression lines in Figure 2.2 are known as the sample regression lines. Supposedly they represent the population regression line but due to sampling fluctuations they are at best an approximation of the true PR. In general, we may get N different SRF's for N different samples and these SRF's are not likely the same.

Like the *PRF* (derived from population regression line) we can develop the concept of **the sample regression function (SRF)** to represent the sample regression line.

Since from the population regression function we know that the function is of the form, $Y = E(Y/X_i) = \alpha + \beta X_i$, the sample counterpart of this equation may be written as

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X_i \text{ where}$$

$\hat{Y}$ = estimator of $E(Y/X_i)$, $\hat{\alpha}$ = estimator of α and $\hat{\beta}$ = estimator of β

It should be noted that an **estimator**, also known as a (sample) statistic, is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator is known as an **estimate**. It should be noted that an estimator is random but an estimate is non random.

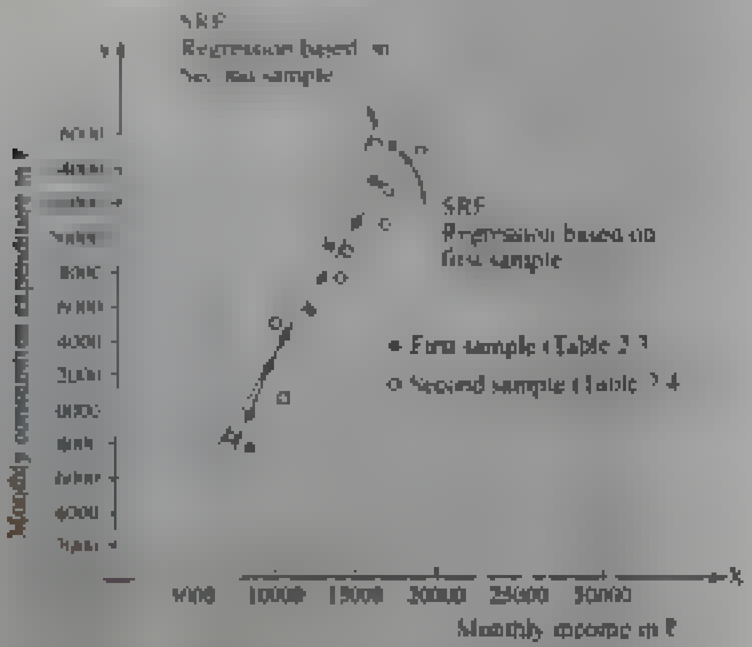


Fig. 2.2. Regression lines based on two different samples.

Like the population regression function ($Y_i = E(Y/X_i) + u_i = \alpha + \beta X_i + u_i$) we can express the sample regression equation $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ in its stochastic form as follows:

$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$ where $\hat{u}_i$ denotes the (sample) residual term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an estimate of u_i. It is introduced in the SRF for the same reasons as u_i was introduced in the *PRF*.

Now our primary objective in regression analysis is to estimate the *PRF* $Y_i = \alpha + \beta X_i + u_i$ on the basis of the *SRF* $Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$. It should be noted that the estimate of the *PRF* based on the *SRF* is at best an approximate one (as sampling fluctuations exist). We have to develop procedures that tell us how to construct the *SRF* to mirror the *PRF* as faithfully as possible.

2.2 The Simple Linear Regression Model

Relationships suggested by economic theory are usually specified as exact or deterministic relationships between variables, while on the other hand much stress is placed on the need for testing these economic theories. This implies a belief in the existence of stochastic function. The knowledge of econometrics tries to test these theoretical propositions in terms of stochastic variables. The simplest form of stochastic *relation* between two variables X and Y is called a simple linear regression model and is given by $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots, n$ where Y = dependent variable, X = explanatory variable (independent variable), u = stochastic disturbance term, α and β are two regression parameters whose values are to be determined on the basis of the given data on X and Y. The subscript i refers to the ith observation, n = sample size or number of data points.

The stochastic nature of the regression model implies that for every value of X there is a whole probability distribution of values of Y. In other words the value of Y can never be predicted exactly. This uncertainty concerning the value of Y arises because of the presence of the stochastic term u, which imparts randomness to Y.

2.2.1 Role of Random disturbance term in Econometric Model

We may ask why should we add an error term or random disturbance term u to an econometric model ?

The disturbance term u is a surrogate for all those variables that are omitted from the model but that collectively affect Y. We may give the following reasons for the insertion of the disturbance term in an econometric model.

(i) ***Omission of variables from the function*** : Suppose in the model $Y = \alpha + \beta X + u$ the variable Y denotes the consumption expenditure and X denotes disposable income. But in reality X is not the only variable influencing Y. The family size, tastes of the family, spending habits and so on affect the variable Y. The error u is a catch-all for the effects of all these variables, some of which may not even be quantifiable and some of which may not even be identifiable. Therefore u may be used as a substitute for all the excluded or omitted variables from the model.

(ii) ***Unpredictable element of randomness in human responses*** : For instance, if Y = consumption expenditure of a household and X = disposable income of the household, there is an unpredictable element of randomness in each household's consumption. The household does not behave like a machine. In one month the people in the household are on a spending spree. In other month they are tightfisted.

(iii) ***Imperfect specification of the mathematical form of the model*** : We may have linearised a possibly nonlinear relationship between Y and X or we may have left out of the model some equations.

It is because the economic phenomena are much more complex than a simple equation may reveal. For example, price determines and is determined by the quantity supplied (or quantity demanded) in the market. Under such circumstances if we attempt to study the phenomena with a single equation model, we are bound to commit an error. Thus the disturbance term represents such an error which may be due to imperfect specification of the form of the model, that is, of the number of equations.

(iv) ***Core variables versus peripheral variables*** : In consumption income relation, for instance, we may observe that besides income X_1, the number of children per family X_2, sex X_3, religion X_4, education X_5 and geographical region X_6 etc. also can affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small that as a practical matter it does not pay to introduce them into the model explicitly. However, their combined effect can be treated as a random variable u_i.

(v) ***Principle of Parsimony*** : Generally we would like to keep our regression model as simple as possible. If we can explain the behaviour of Y substantially with two or three explanatory variables and if our theory is not strong enough to suggest what other variables might be included, why introduce more variables ? In such cases u

[illegible]

Due to aggregation We [illegible] aggregate data [illegible]

Due to errors in measurement (Errors in variables) [illegible] errors in the observations.

For all these reasons the stochastic disturbance term assumes an extremely important role in regression analysis.

2.5 Classical Linear Regression Model and its Assumptions

Let us consider an observed relation between two variables X and Y which is given by $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots, n$ where Y is the dependent variable, X is the independent variable, u is the disturbance term, the subscript i denotes the ith set of observations, α and β are the two parameters whose values are to be estimated on the basis of the observed data on X and Y.

Now the model is called a Classical Linear Regression Model (CLRM) if the model satisfies the following properties (assumptions).

Assumption 1 u_i is a random variable which follows normal distribution.

Assumption 2 $E(u_i) = 0$ for each i, $i = 1, 2, \ldots, n$. This means that the probability distribution of the disturbance term is such that its mean is zero.

Now $E(u_i) = 0$ implies $E(Y_i) = \alpha + \beta X_i$. This can be shown as follows. Since $Y_i = \alpha + \beta X_i + u_i$. Now $E(Y_i) = E(\alpha + \beta X_i + u_i) = E(\alpha) + \beta E(X_i) + E(u_i) = \alpha + \beta X_i$, as $E(u_i) = 0$ and $E(X_i) = X_i$.

But $\alpha + \beta X_i$ is the true value of Y. This means that expectation of observed value of the dependent variable is its true value. In other words the probability distribution of Y_i is centred around the true relationship.

Assumption 3 Variance of each u_i is a constant and is independent of i, $i = 1, 2, \ldots, n$ and is denoted by σ_u^2 or simply σ^2

i.e. $\text{Var}(u_i) = \sigma_u^2$ or σ^2

or, $E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma_u^2$ where $E(u_i) = 0$

Assumptions 2 and 3 imply that the random variables $u_1, u_2, \ldots, u_n$ are identically distributed with the same mean (zero) and same variance (σ_u^2)

i.e. $u_i \sim IID(0, \sigma_u^2)$ for each i, $i = 1, 2, 3, \ldots, n$

Assumption 4 The different error terms are independently distributed, i.e., $E(u_i u_j) = 0$ for $i \neq j$.

Now $\text{Cov}(u_i, u_j) = 0$ for $i \neq j$

and $\text{Cov}(u_i, u_j) = E(u_i u_j)$ [illegible] where [illegible] $\neq j$.

Assumption 5 The independent variable X is non-stochastic [illegible] which [illegible] random [illegible] and [illegible] is independent of explanatory variables.

i.e., $E(X_i u_i) = X_i E(u_i) = 0$ for all $i = 1, \ldots, n$

The regression equation $Y = \alpha + \beta X + u$ along with the given five assumptions represents the Classical Linear Regression Model. The five assumptions play important roles in the sampling distributions of parameters etc.

The effect of first three assumptions on the probability distribution of dependent variable Y can now be rationalised.

(i) In the equation $Y = \alpha + \beta X + u$, Y is a linear function of u. Since u is normally distributed, it follows that Y is also normally distributed.

(ii) $Y_i = \alpha + \beta X_i + u_i$

$$E(Y_i) = E(\alpha + \beta X_i + u_i) \qquad [\because E(u_i) = 0]$$

$$= \alpha + \beta X_i$$

This means that the mean of Y_i is $\alpha + \beta X_i$

(iii) $\text{Var}(Y_i) = E[Y_i - E(Y_i)]^2 = E[\alpha + \beta X_i + u_i - \alpha - \beta X_i]^2$

$$= E(u_i)^2 = \sigma_u^2 \left[\because E(u_i)^2 = \sigma_u^2\right]$$

Therefore, we say that variance of Y_i is σ_u^2.

Thus with the first three assumptions of u, we can indirectly say that Y is normally distributed with mean $(\alpha + \beta X_i)$ and variance σ_u^2.

Symbolically, $Y_i \sim N(\alpha + \beta X_i, \sigma_u^2)$ when $u_i \sim N(0, \sigma_u^2)$. This is illustrated in Fig. 2.3. Let $Y = \alpha + \beta X$ represent the population regression line. This regression line is unknown as we do not know the exact values of α and β. We have to estimate the values of α and β on the basis of sample data.

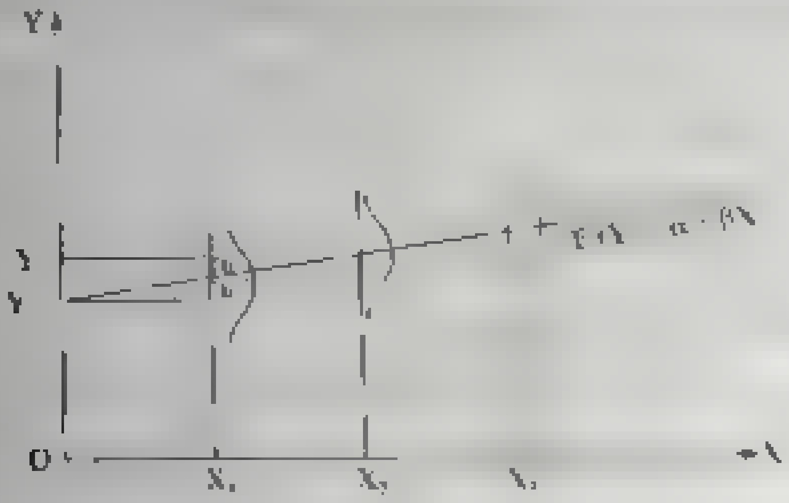


Fig. 2.3.

[illegible]

[illegible] α and β respectively then [illegible] represents the fitted values [illegible] sample regression line [illegible]

[illegible] and the difference [illegible] residual [illegible] Then the true relationship is given by [illegible] and the estimated relationship given by [illegible] The difference between these two results along with the difference between [illegible] has been illustrated in Fig. 2.4

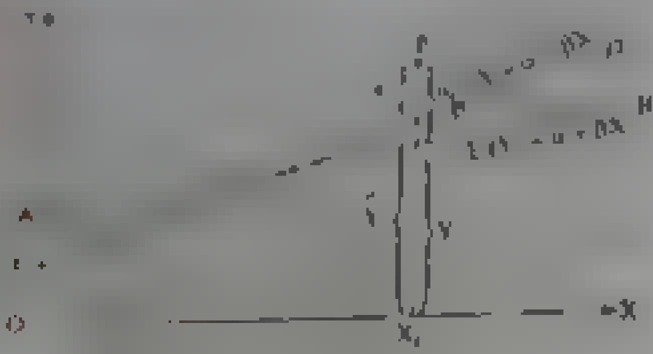


Fig. 2.4.

In Fig. 2.4, *AB* is the true regression line. *CD* is the estimated regression line [illegible] represents one of the observations in the sample data. *e* differs from *u* because the true values of the parameters are different from their estimated values. In fact, we can think the residual *e* as the estimate of the disturbance u_i

2.4 Methods of Estimating Regression Parameters

There are different methods for estimating the regression parameters.

Now we shall discuss three methods for estimating the regression parameters α and β. These are

(i) The method of moments

(ii) The method of ordinary least squares (OLS)

(iii) The method of maximum likelihood (MLE)

2.5 The Method of Moments

The assumptions we have made (Assumption-4) about the error term *u* imply that $E(u) = 0$ and $Cov(X, u) = 0$.

In the method of moments, we replace these conditions by their sample counterparts.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimators of α and β respectively.

Since $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ [illegible] is the regression equation.

The sample counterpart of u_i is the estimated error $\hat{u}_i$ which is also called the residual, defined as $\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$.

The two equations to determine $\hat{\alpha}$ and $\hat{\beta}$ are obtained by replacing population assumptions by their sample counterparts.

Population Assumption	**Sample Counterpart**
$E(u) = 0$	$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$ or $\sum_{i=1}^{n} \hat{u}_i = 0$
$\text{cov}(X, u) = 0$	$\frac{1}{n}\sum_{i=1}^{n} X_i\hat{u}_i = 0$ or $\sum_{i=1}^{n} X_i\hat{u}_i = 0$

Thus we get the two equations

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ or } \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

and

$$\sum_{i=1}^{n} X_i\hat{u}_i = 0 \text{ or } \sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

These equations can be written as

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i \qquad \text{......(1)}$$

$$\sum_{i=1}^{n} X_iY_i = \hat{\alpha}\sum X_i + \hat{\beta}\sum X_i^2 \qquad \text{......(2)}$$

These two equations are called "normal equations". Solving these two equations we can get $\hat{\alpha}$ and $\hat{\beta}$.

Example 2.1. Consider the data on advertising expenditures (X) and sales revenue (Y) for an athletic sports wear store for 5 months.

The observations are as follows

Month	Sales Revenue (Y) (in 000 ₹)	Advertising Expenditure (X) (in 00 ₹)
1	3	1
2	4	2
3	2	3
4	6	4
5	8	5

Solution. Let $Y_i = \alpha + \beta X_i + u_i$ be the regression equation. The two normal equations for estimating the regression coefficients are

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i \qquad (1)$$

$$\sum_{i=1}^{n} X_iY_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2 \qquad (2)$$

Calculations for $\hat{\alpha}$ and $\hat{\beta}$ (estimators of α & β)

Month	X	Y	X^2	XY	$\hat{u}$
1	1	3	1	3	[illegible]
2	2	4	4	8	[illegible]
3	3	[illegible]	9	[illegible]	[illegible]
4	4	6	16	[illegible]	[illegible]
5	5	[illegible]	25	[illegible]	[illegible]
Total	$\sum X = 15$	$\sum Y =$ [illegible]	$\sum X^2 =$ [illegible]	$\sum XY =$ [illegible]	$\sum \hat{u} =$ [illegible]

where n = 5. Now putting in the two normal equations

we get $5\alpha + 15\beta =$ [illegible] (1)

$15\alpha + 55\beta =$ [illegible] (2)

Solving (1) and (2) by Cramer's rule we get

$\hat{\alpha} =$ [illegible] and $\hat{\beta} =$ [illegible]

Thus the estimated regression equation is

$\hat{Y} =$ [illegible]

The intercept $\hat{\alpha}$ gives the value of Y when X = 0. This says that [illegible] expenditures are zero, sales revenue will be ₹ [illegible]. The slope coefficient [illegible] says that if advertisement expenditure (X) is changed by [illegible] unit, [illegible] increases by [illegible] units (₹ [illegible]) on an average. We have also shown the estimated values of the residuals, given by

$\hat{u}_i = Y_i - \hat{Y}_i$ [illegible] shown in the last column of the above table.

2.6. The Method of Ordinary Least Squares (OLS)

Let $Y_i = \alpha + \beta X_i + u_i$ be a two-variable linear regression model, where Y is the dependent variable and X is the independent variable and u is the disturbance term. If the disturbance term u satisfies the following properties, then this model will be called a classical linear regression model (CLRM):

(i) $E(u_i) = 0$ for each i = 1, 2, 3, …, n

(ii) $V(u_i) = \sigma_u^2$ for each i

(iii) $E(u_i u_j) = 0$ for all i ≠ j

(iv) $E(u_i u_j) = \sigma_u^2$ for all i = j

(v) Independent variable X is non-stochastic.

The two parameters α and β of the regression equation can be obtained by the method of ordinary least squares (OLS). Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimated values of and β. The estimated relation becomes $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ and $e_i = Y_i - \hat{Y}_i$ is the residual term which shows the difference between the observed and estimated value.

The method of least squares consists in finding out those values of α and β for which $\sum_{i=1}^{n} e_i^2$ is minimum. This means that we have to minimise $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$ through the choice of α and β. The necessary conditions of minimisation require

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \alpha} = -2\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i) = 0 \qquad (1)$$

and

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta} = -2\sum_{i=1}^{n} X_i (Y_i - \alpha - \beta X_i) = 0 \qquad (2)$$

Simplifying equations (1) and (2) we get two normal equations

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} X_i \qquad (3)$$

$$\sum_{i=1}^{n} X_i Y_i = \alpha \sum_{i=1}^{n} X_i + \beta \sum_{i=1}^{n} X_i^2 \qquad (4)$$

Now solving equations (3) and (4) by Cramer's rule we have

$$\hat{\beta} = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad ([illegible])$$

or $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{assuming } x_i = X_i - \bar{X} \text{ and } y_i = Y_i - \bar{Y} =$$

Again from equation (3) we get

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta\sum_{i=1}^{n} X_i$$

or, $$\sum_{i=1}^{n}\frac{Y_i}{n} = \alpha + \beta\sum_{i=1}^{n}\frac{X_i}{n} \quad \text{or} \quad \bar{Y} = \alpha + \beta\bar{X} \qquad \therefore \alpha = \bar{Y} - \beta\bar{X}$$

2.6.1 Reverse Regression

By applying OLS method we have estimated the linear regression equation $Y_i = \alpha + \beta X_i + u_i$ where u_i satisfies all the properties of CLRM. The estimated regression equation becomes $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of α and β.

$$\text{Here } \hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)} = \frac{r_{XY}\sigma_X\sigma_Y}{\sigma_X^2} = r_{XY}\frac{\sigma_Y}{\sigma_X}$$

If we put $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ then we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Here $\hat{\beta}$ is the estimated regression coefficient of Y on X. In this case the regression equation so defined is called the direct regression equation (of Y on X) where Y is the dependent variable and X is the independent variable. Sometimes we have to consider the regression equation of X on Y as well. This is called **reverse regression.**

The reverse regression is used in many cases. For instance, reverse regression has been advocated in the analysis of sex (or race) discrimination in salaries.

Suppose Y = salary and X = qualification and we are interested in determining if there is sex discrimination in salaries. We can ask

1. Whether men and women with the same qualifications (value of X) are getting the same salaries (value of Y). This question is answered by the direct regression, i.e., regression of Y on X. Alternatively, we can ask
2. Whether men and women with same salaries (value of Y) have the same qualifications (value of X).

This question is answered by the reverse regression, i.e., regression equation of X on Y.

For the reverse regression, the regression equation can be written as $X_i = \alpha' + \beta' Y_i + v_i$ where v_i are the errors satisfying all the properties of CLRM. Here X is the dependent variable and Y is the independent variable.

The estimated relation becomes $\hat{X}_i = \hat{\alpha} + \hat{\beta} Y_i$ and $e_i = X_i - \hat{X}_i$ which is the residual term showing the difference between the observed and estimated relation. The method of least squares consists in finding out those values of $\hat{\alpha}$ and $\hat{\beta}$ for which $\sum_{i=1}^{n} e_i^2$ is minimum. This means that we have to minimise

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (X_i - \hat{X}_i)^2 = \sum_{i=1}^{n} (X_i - \hat{\alpha} - \hat{\beta} Y_i)^2$ through the choice of $\hat{\alpha}$ and $\hat{\beta}$. The necessary conditions of minimisation require

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n} (X_i - \hat{\alpha} - \hat{\beta} Y_i) = 0 \qquad \ldots (1)$$

and
$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = -2\sum_{i=1}^{n} Y_i (X_i - \hat{\alpha} - \hat{\beta} Y_i) = 0 \qquad \ldots (2)$$

Simplifying equations (1) and (2) we get two normal equations

$$\sum_{i=1}^{n} X_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} Y_i \qquad \ldots (3)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} Y_i + \hat{\beta} \sum_{i=1}^{n} Y_i^2 \qquad \ldots (4)$$

Now solving equations (3) and (4) by Cramer's rule we have,

$$\hat{\beta} = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} X_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} Y_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}$$

$$= \frac{Cov(X, Y)}{Var(Y)} = \frac{r_{XY}\sigma_X\sigma_Y}{\sigma_Y^2} = r_{XY}\frac{\sigma_X}{\sigma_Y}$$

= Regression coefficient of X on Y

If we [illegible] then the [illegible] as follow

[illegible]

Again [illegible]

[illegible]

or [illegible] $\beta \bar{X}$

It should be noted that β is the regression coefficient of Y on X and [illegible] regression [illegible] of X on Y.

Since $\beta = r_{XY}\frac{\sigma_Y}{\sigma_X}$ and $\beta' = r_{XY}\frac{\sigma_X}{\sigma_Y}$

$$\beta\beta' = r_{XY}\frac{\sigma_Y}{\sigma_X}\, r_{XY}\frac{\sigma_X}{\sigma_Y} = r^2_{XY}$$

The two regression lines (Y on X and X on Y) will be different [illegible] r_{XY} [illegible]. The two regression lines will coincide if $r_{XY} =$ [illegible] and they will [illegible] perpendicular to each other if $r_{XY} = 0$.

Example 2.1.1. We now consider a numerical example where we will find both the regression (direct regression i.e. Y on X and reverse regression i.e. X on Y) on the basis of the following data

X	10	7	10	5	8	8	6	7	9	10
Y	11	10	12	6	10	7	9	10	[illegible]	11

where X = labour-hours of work, Y = output

Solution We know that the fitted direct regression equation (Y on X) is given by

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X \quad \text{where } \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. Conversely, the equation of the fitted reverse regression equation (X on Y) is given by

$$\hat{X} = \hat{\alpha}' + \hat{\beta}' Y \quad \text{where } \hat{\beta}' = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2}, \text{ where } \hat{\alpha}' = \bar{X} - \hat{\beta}'\bar{Y} \text{ and } x_i = X_i - \bar{X}$$

Calculations for direct and reverse regression lines

Observation	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4	[illegible]
2	7	10	−1	0.4	−0.4	1	[illegible]
3	10	12	2	2.4	4.8	4	5.76
4	[illegible]	8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	8	10	0	0.4	0	0	0.16
6	8	7	0	−2.6	0	0	6.76
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	7	10	−1	0.4	−0.4	[illegible]	[illegible]
9	9	11	1	1.4	1.4	1	1.96
10	10	10	[illegible]	0.4	[illegible]	4	[illegible]
Total	80	96	0	0	[illegible]	[illegible]	30.4

$\bar{X} = \frac{\Sigma X}{n} = \frac{80}{10} = 8$, $\bar{Y} = \frac{\Sigma Y}{n} = \frac{96}{10} = 9.6$

Now $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{21}{28} = 0.75$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 9.6 - 0.75 \times 8 = 3.6$

Estimated regression equation of Y on X (direct regression) is given by $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 3.6 + 0.75\,X$

Again, $\hat{\beta}' = \frac{\Sigma xy}{\Sigma y^2} = \frac{21}{30.4} = 0.690$

and $\hat{\alpha}' = \bar{X} - \hat{\beta}'\bar{Y} = 8 - 0.690 \times 9.6 = 1.376$

Estimated reverse regression equation (X on Y) is given by $\hat{X} = \hat{\alpha}' + \hat{\beta}' Y$

or, $\hat{X} = 1.376 + 0.690\,Y$

It should be noted that $\hat{\beta} = 0.75$ is the estimated regression coefficient of Y on X and $\hat{\beta}' = 0.690$ is the estimated regression coefficient of X on Y.

Since $\hat{\beta} \times \hat{\beta}' = r_{XY}^2$

$r_{XY}^2 = 0.75 \times 0.690 = 0.5175 = 0.52$ and $r_{XY} = \sqrt{0.5175} = 0.72$

2.6.2. Scaling and Units of Measurement

In the regression analysis the units in which the regressand or the dependent variable (Y) and the regressor(s) are measured make difference in the regression results.

Suppose we like to regress Indian gross domestic savings (GDS) and gross domestic product (GDP) in rupees crore as well in rupees lakh measured in 1999-2000 prices. We us assume that in the regression of GDS on GDP one researcher uses data in rupees crore and another researcher uses data in rupees lakh. Now the natural question is: will the regression results be the same in both cases? Do the units in which the regressand and regressor(s) are measured make any difference in the regression results? If so, what is the sensible course to follow in choosing units of measurement for regression analysis? To answer these questions, let us proceed as follows.

Let $Y_i = \alpha + \beta X_i + u_i$ (1)

where Y = GDS and X = GDP. Let us define

$Y_i^* = w_1 Y_i$ (2)

and $X_i^* = w_2 X_i$ (3)

where w_1 and w_2 are constants, called the scale factors. w_1 may be equal to w_2 or may be different. If Y_i and X_i are measured in say rupees crore and we want to express them in rupees lakh, we will have $Y_i^* = 100\,Y_i$ and $X_i^* = 100\,X_i$; here $w_1 = w_2 = 100$.

Now consider the regression using Y_i^* and X_i^* variables

$Y_i^* = \alpha^* + \beta^* X_i^* + u_i^*$ (4)

where $Y_i^* = w_1 Y_i$, $X_i^* = w_2 X_i$ and $u_i^* = w_1 u_i$, as $w_1 = w_2$

Now comparing equations (1) and (4) we can find out the relationships between the following pairs:

1. $\hat{\alpha}$ and $\hat{\alpha}^*$
2. $\hat{\beta}$ and $\hat{\beta}^*$
3. Var ($\hat{\alpha}$) and Var ($\hat{\alpha}^*$)
4. Var ($\hat{\beta}$) and Var ($\hat{\beta}^*$)
5. $\hat{\sigma}_u^2$ and $\hat{\sigma}_u^{2*}$
6. r^2_{XY} and $r^2_{X^*Y^*}$

From the least squares theory we know that (applying OLS method on equation 1)

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ (5)

$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ (6) where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$

Var ($\hat{\alpha}$) $= \frac{\sum X_i^2}{n\sum x_i^2}\cdot \sigma_u^2$ (7)

Var ($\hat{\beta}$) $= \frac{\sigma_u^2}{\sum x_i^2}$ (8)

and $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$ or $\frac{\sum e_i^2}{n-2}$ (9)

Similarly applying OLS method to equation (4) we obtain

$$\hat{\alpha}^* = \bar{Y}^* - \hat{\beta}^* \bar{X}^* \qquad \text{—— (10)}$$

$$\hat{\beta}^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} \qquad \text{(11) where } x_i^* = X_i^* - \bar{X}^* \text{ and } y_i^* = Y_i^* - \bar{Y}^*$$

$$\text{Var}(\hat{\alpha}^*) = \frac{\sum X_i^{*2}}{n \sum x_i^{*2}} \cdot \sigma_u^{*2} \qquad \text{(12)}$$

$$\text{Var}(\hat{\beta}^*) = \frac{\sigma_u^{*2}}{\sum x_i^{*2}} \qquad \text{(13)}$$

$$\hat{\sigma}_u^{*2} = \frac{\sum \hat{u}_i^{*2}}{n-2} \text{ or } \frac{\sum e_i^{*2}}{n-2} \qquad \text{(14)}$$

Thus we see that from model (1) $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of α and β and from model (2) $\hat{\alpha}^*$ and $\hat{\beta}^*$ are the OLS estimators of α^* and β^*. From the above results it is easy to establish relationship between two sets of parameters.

Since $Y^* = W_1 Y$ (or $y_i^* = w_1 y_i$), $X_i^* = W_2 X_i$ (or $x_i^* = w_2 x_i$), $\hat{u}_i^* = w_1 \hat{u}_i$, $\bar{Y}^* = W_1 \bar{Y}$ and $\bar{X}^* = W_2 \bar{X}$, we can easily verify that

$$\hat{\beta}_i^* = \left(\frac{W_1}{W_2}\right)\hat{\beta} \qquad \text{—— (15)}$$

$$\hat{\alpha}^* = W_1 \hat{\alpha} \qquad \text{—— (16)}$$

$$\hat{\sigma}_u^{*2} = W_1^2 \hat{\sigma}_u^2 \qquad \text{(17)}$$

$$\text{Var}(\hat{\alpha}^*) = W_1^2 \ \text{Var}(\hat{\alpha}) \qquad \text{—— (18)}$$

$$\text{Var}(\hat{\beta}^*) = \left(\frac{W_1}{W_2}\right)^2 \text{Var}(\hat{\beta}) \qquad \text{(19)}$$

$$r_{XY}^2 = r_{X^*Y^*}^2 \qquad \text{(20)}$$

From the above results it is clear that from the regression results based on one scale of measurement, we can derive the results based on another scale of measurement once the scaling factors are known. From the results given in (15) to (20) we can also derive some special cases. For instance, if the scaling factors are identical (i.e. $W_1 = W_2$) the slope coefficient and its standard error remain unaffected in going from the (Y_i, X_i) to the (Y_i^*, X_i^*) scale. However, the intercept and its standard error are both multiplied by W_1 (when $W_1 = W_2$). But if the X scale is not changed (i.e. $W_2 = 1$) and the Y scale is changed by the factor W_1, the slope as well as the intercept coefficients and their respective standard errors are all multiplied by the same W_1 factor. Finally, if the Y scale remains unchanged [illegible] but the X scale is changed by the factor [illegible] the slope coefficient and its standard error are multiplied by the factor [illegible] but the intercept coefficient and the standard error remain unaffected.

It should, however, be noted that the transformation [illegible] variables from the [illegible] the (X^*, Y^*) scale does not affect the properties of OLS estimators.

To illustrate the above theoretical results we consider an example showing the relationship between GDS and GDP of India during the period [illegible] to [illegible]. The estimated regression equation of GDS on GDP (both GDS and GDP in rupees crore) is given by

$$\widehat{GDS}_t = -6742.37 + 0.36\,GDP_t \qquad \ldots(2.1)$$
$$SE \qquad (7721.01) \quad (0.02) \qquad r^2 = 0.8891$$

Similarly, the estimated regression equation of GDS on GDP (both GDS and GDP in rupees lakh) is given by

$$\widehat{GDS}_t = -674236.51 + 0.36\,GDP_t \qquad \ldots(2.2)$$
$$SE \qquad (772100.74) \quad (0.02) \qquad r^2 = 0.8891$$

Here we see that the intercept and its standard error is 100 times the corresponding values in the regression (2.1) (we should note that $w_1 = 100$ in going from crore to lakhs of rupees, i.e., 1 crore = 100 lakhs), but the slope coefficient as well as its standard error is unchanged, in accordance with the theory.

Now suppose we measure GDS in rupees crore and GDP in rupees lakh. The estimated regression equation becomes

$$\widehat{GDS}_t = -6742.37 + 0.0036\,GDP_t \qquad \ldots(2.3)$$
$$SE \qquad (7721.01) \quad (0.0002) \qquad r^2 = 0.8891$$

As expected, the slope coefficient as well as the standard error is [illegible] the value in equation (2.1) since only X or GDP scale is changed.

If we express GDS in rupees lakh and GDP in rupees crore, the estimated regression equation becomes

$$\widehat{GDS}_t = -674236.51 + 36.33\,GDP_t \qquad \ldots(2.4)$$
$$SE \qquad (772100.74) \quad (1.78) \qquad r^2 = 0.8891$$

Here we see that both the intercept and the slope coefficients as well as their respective standard errors are 100 times their values in equation (2.1), in accordance with our theoretical results.

It should be noted that the r^2 value remains the same in all the cases as it is invariant to changes in the unit of measurement and scales.

2.7 Estimation of a Function whose Intercept is Zero

In some cases economic theory postulates relationships which have a zero intercept, that is, they pass through the origin of the XY plane. [For example, long run consumption function of the form $C = bY$ where b = APC = MPC, C = consumption expenditure, Y = income].

In this event we should estimate the function $Y = \alpha + \beta X + u$, imposing the restriction $\alpha = 0$. The formula for the estimation of β then becomes $\hat{\beta} = \dfrac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$ which involves the actual values of the variables, and not their deviations, as in the case of unrestricted value of α.

Proof We want to fit the line $Y_i = \alpha + \beta X_i + u_i$, subject to the restriction $\alpha = 0$. To estimate $\hat{\beta}$, the problem is put in a form of restricted minimization problem and then Lagrange method is applied.

Now we have to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

subject to $\hat{\alpha} = 0$

The Lagrange composite function then becomes $L = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 - \lambda\hat{\alpha}$, where λ is the Lagrange multiplier. Now we have to minimize L with respect to $\hat{\alpha}$, $\hat{\beta}$ and λ. First order conditions of minimization require

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i) - \lambda = 0 \qquad (1)$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0 \qquad (2)$$

$$\frac{\partial L}{\partial \lambda} = -\hat{\alpha} = 0 \qquad (3)$$

Now substituting (3) in (2) and rearranging we get

$$-2\sum_{i=1}^{n} X_i (Y_i - \hat{\beta} X_i) = 0$$

$$\text{or } \sum_{i=1}^{n} X_i Y_i - \hat{\beta}\sum_{i=1}^{n} X_i^2 = 0 \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

In this case (i) $\hat{\sigma}_u^2 = \Sigma e_i^2$ [illegible] (ii) [illegible] $SE(\hat{\beta}) = \sqrt{\sigma_u^2}$ [illegible] $R^2 = $ [illegible] Σe_i^2 [illegible] ΣY^2

2.6 Estimation of Elasticities from an Estimated Regression Line

The estimated regression equation is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ whose [illegible] are [illegible]. $\hat{\beta}$ is the derivative of $\hat{Y}$ with respect to X, i.e. $\hat{\beta} = \frac{dY}{dX}$ which shows the rate of change in Y as X changes by a very small amount. It should be clear that the estimated function is a linear demand or supply function, the coefficient $\hat{\beta}$ is not the price elasticity, but a component of the elasticity which is defined by the formula

$$\eta_p = \frac{dY}{dX}\cdot\frac{X}{Y} = \frac{X}{Y}\cdot\frac{dY}{dX}$$

where η_p = price elasticity, Y = quantity (demanded or supplied), X = price and $\hat{\beta}$ is the component $\frac{dY}{dX}$. From an estimated function we can obtain an average elasticity $\eta_p = \hat{\beta}\frac{\bar{X}}{\hat{\bar{Y}}}$ where $\bar{X}$ = the average price in the sample, $\hat{\bar{Y}}$ = average regressed value of the quantity, i.e. the mean value as estimated from the regression $\hat{Y}$. It should be noted that $\hat{\bar{Y}} = \bar{Y}$, the average value of the quantity in the sample.

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$

$$\hat{\bar{Y}} = \hat{\alpha} + \hat{\beta}\bar{X} = \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}\bar{X} = \bar{Y}$$

In particular if $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, is the regression equation, then the estimated average elasticity $\eta_p = \hat{\beta}\frac{\bar{X}}{\hat{\bar{Y}}}$ where $\hat{\bar{Y}} = \hat{\alpha} + \hat{\beta}\bar{X}$

Now substituting for $\hat{\bar{Y}}$ in the expression of elasticities, we obtain $\eta_p = \frac{\hat{\beta}\bar{X}}{\hat{\alpha} + \hat{\beta}\bar{X}}$

If the function $Y = \alpha + \beta X$ represents a supply function with $\beta > 0$, it follows that

(i) the supply function will be elastic ($\eta_p > 1$) if α is negative ($\alpha < 0$)

(ii) the supply function will be inelastic ($\eta_p < 1$) if $\alpha > 0$

(iii) the supply function will have unitary elasticity ($\eta_p = 1$) if $\alpha = 0$

Thus the elasticity of a supply curve (with positive slope) depends on the sign of the constant intercept α.

Example 2.2 The following table includes the price and quantity demanded of the product of a monopolist over a six year period.

Year	2014	2015	2016	2017	2018	2019
Quantity (000 Kg.)	5	3	4	7	8	0
Price (00 ₹)	2	4	3	1	3	5

(a) Estimate the demand function assuming a linear demand function. Comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

(b) Estimate the average elasticity of demand.

(c) Estimate the elasticity of demand at the price 4.

(d) Forecast the level of demand if price rises to 5. Comment on your forecast.

Solution (a) Let $Y_t = \alpha + \beta X_t$ for $t = 1, 2, \ldots, 6$ be the linear demand function. By the OLS method we can get the estimators of α and β. Here Y = demand, X = price, α, β are two parameters. Theoretically we may assume $\alpha > 0$, $\beta < 0$. By OLS method $\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2}$ where $x_t = X_t - \bar{X}$, $y_t = Y_t - \bar{Y}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, $\bar{X} = \sum X_t / n$, $\bar{Y} = \sum Y_t / n$

Calculations for the parameters (α, β)

Year (t)	quantity ('000 kg) Y_t	price ('00 ₹) X_t	$y_t = Y_t - \bar{Y}$	$x_t = X_t - \bar{X}$	$x_t y_t$	x_t^2
2014 (1)	8	2	3	1	3	
2015 (2)	3	4	2	1	2	1
2016 (3)	4	3	−1	0	0	0
2017 (4)	7	1	−2	2	4	4
2018 (5)	8	3	3	0	0	0
2019 (6)	0	5	−5	2	10	4
n = 6 Total	$\sum Y_t = 30$	$\sum X_t = 18$	$\sum y_t = 0$	$\sum x_t = 0$	$\sum x_t y_t = -19$	$\sum x_t^2 = 10$

$$\bar{Y} = \frac{\sum Y_t}{n} = \frac{30}{6} = 5, \quad \bar{X} = \frac{\sum X_t}{n} = \frac{18}{6} = 3$$

Now $$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{-19}{10} = -1.9$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 5 - (-1.9) \times 3 = 5 + 5.7 = 10.7$

Thus the OLS estimators of α and β are $\hat{\alpha} = 10.7 > 0$ and $\hat{\beta} = -1.9 < 0$.

Therefore the estimated demand function is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $Y = \hat{Y} = 10.7 - 1.9X$

This is consistent with the theory where we assume $\alpha > 0$ and $\beta < 0$. This clearly shows that there exists an inverse relation between price and demand, i.e., the law of demand holds true.

(b) The average elasticity (price elasticity of demand) is given by

$$\eta_p = \hat{\beta}\,\frac{\bar{X}}{\bar{Y}} = -1.9 \times \frac{3}{5} = -1.14 \text{ or } |\eta_p| = 1.14 > 1$$

This means that the demand function shows an elastic demand.

(c) We have to estimate η_p (price elasticity of demand), from the estimated relation $\hat{Y} = Y = 10.7 - 1.9X$ when price = $X = 4$

If $X = 4$, $\hat{Y} = 10.7 - 1.9 \times 4 = 10.7 - 7.6 = 3.1$

$$\text{Now } \eta_p \text{ (at } X = 4) = \frac{X}{Y}\,\frac{dY}{dX} = \frac{4}{3.1} \times (-1.9) = -2.45$$

$|\eta_p|$ at $X = 4$ is $2.45 > 1$

This implies that the demand is elastic demand.

[illegible] demand when price rises [illegible] where [illegible]

[illegible]

[illegible]

This means that [illegible] from [illegible] demand decreases [illegible]

[illegible] price [illegible] demand decreases [illegible]

Example 2.3 The following table shows [illegible] pairs of observations on [illegible] quantity supplied:

No. of observations (n)	1	2	3	4	5	6	7	8	9	10	11	12
Quantity (in tonnes)	69	76	52	56	57	77	58	55	67	53	72	64
Price (in '00 ₹)	9	12	6	10	9	10	7	8	12	6	11	8

(a) Assuming a linear supply function estimate the supply function [illegible] comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

(b) Estimate the average price elasticity of supply.

(c) Estimate the elasticity of supply at the price [illegible]

(d) Forecast the level of supply [illegible] price rises to ₹ [illegible]

Solution Let $Y = \alpha + \beta X + u$ for $i = 1, \ldots, 12$ be the linear supply function. [illegible] Y = supply, X = price [illegible] method we can get the estimators of α and β. Here [illegible] By OLS method we may get [illegible] parameters. Theoretically we may assume α [illegible] 0 and β [illegible] 0.

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{where} \quad x_i = X_i - \bar{X},\ y_i = Y_i - \bar{Y}$$

$$\text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}, \quad \bar{Y} = \frac{\sum Y_i}{n}, \quad \bar{X} = \frac{\sum X_i}{n}$$

Calculations for the OLS estimators of parameters (α, β)

Observations n	Y_i Quantity (in tonnes)	X_i Price (in '00 ₹)	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	$x_i y_i$	x_i^2
1	69	9	0	6	0	0
2	76	12	3	13	39	9
3	52	6	−3	−11	33	9
4	56	10	1	−7	−7	1
5	57	9	0	−6	0	0
$n = 12$ 6	77	10	1	14	14	1
7	58	7	−2	−5	10	4
8	55	8	−1	−8	8	1
9	67	12	3	4	12	9
10	53	6	−3	−10	30	9
11	72	11	2	9	18	4
12	64	8	−1	1	−1	1
Total	$\sum Y_i = 756$	$\sum X_i = 108$	$\sum x_i = 0$	$\sum y_i = 0$	$\sum x_i y_i = 156$	$\sum x_i^2 = 48$

$$\bar{X} = \frac{\sum X}{n} = \frac{108}{12} = 9, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{756}{12} = 63$$

∴ Now the OLS estimators of the regression parameters α and β are given by

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{156}{48} = 3.25$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 63 - 3.25 \times 9 = 63 - 29.25 = 33.75$

Thus the estimated supply function is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $\hat{Y} = 33.75 + 3.25X$

Here we see that $\hat{\alpha} = 33.75 > 0$ and $\hat{\beta} = 3.25 > 0$. This means that there is a direct (positive) relation between supply and price. The intercept of the supply function is positive here. Hence our results are consistent with the theory.

(b) Average price elasticity of supply is given by $\eta_p = \hat{\beta}\,\frac{\bar{X}}{\bar{Y}} = 3.25 \times \frac{9}{63} = .46$

This shows that at the average price the supply is price inelastic.

(c) We have to find price elasticity of supply at price 6.

Since the estimated supply function is

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $\hat{Y} = 33.75 + 3.25X$

Now, $\hat{Y}(X = 6) = 33.75 + 3.25 \times 6 = 33.75 + 19.5 = 53.25$

Now by definition price elasticity of supply $\eta_p = \frac{X}{Y} \cdot \frac{dY}{dX} = \frac{6}{53.25} \times 3.25 = .366$

Thus $\eta_p = 0.366$ when $X = 6$.

(d) From the estimated supply function we see that $\hat{Y} = 33.75 + 3.25X$

When $X = 6$, $\hat{Y} = 53.25$

If now price increases to 8 i.e., if $X = 8$

then $\hat{Y} = 33.75 + 3.25 \times 8 = 33.75 + 26 = 59.75$

This means that when $X = 6$, $\hat{Y} = 53.25$

and when $X = 8$, $\hat{Y} = 59.75$

Thus we may forecast that as price increases, supply will also increase.

2.9 Properties of Least Squares Estimators

The least squares estimates are called BLUE (best linear unbiased estimates) provided that the random term u satisfies some general assumptions, namely that the u has zero mean and constant variance. This proposition, together with the set of conditions under which it is true, is known as **Gauss-Markov Least-Squares Theorem.**

The OLS estimators possess three properties. They are linear, unbiased and have the smallest variance (compared to other linear unbiased estimators). Thus the OLS estimators are BLUE.

1. The property of linearity

The least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the observed sample values Y.

Since $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

Now $\hat{\beta} = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

$$= \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X}) - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \left[\sum_{i=1}^{n}(X_i - \bar{X}) = 0\right]$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{where } x_i = (X_i - \bar{X})$$

Let us suppose that $\dfrac{x_i}{\sum_{i=1}^{n} x_i^2} = K_i \,(i = 1, 2, \ldots, n)$

$$\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$$

This shows that $\hat{\beta}$ is a linear function of Y_i.

Similarly, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i - \bar{X}\sum_{i=1}^{n} K_i Y_i$ $\left[\text{where } \hat{\beta} = \sum_{i=1}^{n} K\right.$

$$\hat{\alpha} = \sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}K_i\right]Y_i$$

This shows that $\hat{\alpha}$ is a linear function of Y_i.

Thus both $\hat{\alpha}$ and $\hat{\beta}$ are expressed as linear functions of the Y's

2 The property of unbiasedness

The means of $\hat\alpha$ $[E(\hat\alpha)]$ and $\hat\beta$ $[E(\hat\beta)]$ can be obtained as follows

Since $$\hat\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n}(X_i - \bar X)(Y_i - \bar Y)}{\sum_{i=1}^{n}(X_i - \bar X)^2}$$

$$= \frac{\sum_{i=1}^{n} Y_i(X_i - \bar X) - \bar Y\sum_{i=1}^{n}(X_i - \bar X)}{\sum_{i=1}^{n}(X_i - \bar X)^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \text{ where } \sum_{i=1}^{n}(X_i - \bar X) = 0$$

and $x_i = X_i - \bar X$ for $i = 1, 2, \ldots, n$

$$\hat\beta = \sum_{i=1}^{n} K_i Y_i \text{ where } K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \text{ and } \hat\alpha = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar X K_i\right) Y_i$$

Now $\hat\beta = \sum_{i=1}^{n} K_i Y_i$. We now put $Y_i = \alpha + \beta X_i + u_i$

$$\hat\beta = \sum_{i=1}^{n} K_i(\alpha + \beta X_i + u_i) = \alpha\sum_{i=1}^{n} K_i + \beta\sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} K_i u_i$$

Since $K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$, $\sum_{i=1}^{n} K_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0$, as $\sum_{i=1}^{n} x_i = 0$

and $\sum_{i=1}^{n} K_i X_i = \sum_{i=1}^{n} K_i(x_i + \bar X)$ where $x_i = X_i - \bar X$ ∴ $X_i = x_i + \bar X$

$$= \sum_{i=1}^{n} K_i x_i + \bar X\sum_{i=1}^{n} K_i = \sum_{i=1}^{n} K_i x_i \quad \left[\because \sum_{i=1}^{n} K_i = 0\right]$$

Now $\sum_{i=1}^{n} K_i x_i = \sum_{i=1}^{n} x_i x_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = 1$ as $K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$

If we put $\sum k_i$ [illegible] and $\sum k_i$ [illegible] in the expression of

$\hat{\beta}$ [illegible] $\sum_{i=1}^{n} k_i + \sum k_i$ [illegible] $\sum k_i u_i$ we get $\hat{\beta} = \beta + \sum_{i=1}^{n} k_i u_i$

Now mean of $\hat{\beta} = E(\hat{\beta}) = E(\beta) + \sum_{i=1}^{n} k_i E(u_i) = \beta$ [illegible]

Thus we have $E(\hat{\beta}) = \beta$ i.e. mean of $\hat{\beta}$ is β

Similarly $\hat{\alpha} = \sum_{i=1}^{n}$ [illegible] $\bar{X}k_i$ [illegible]

$$\sum_{i=1}^{n} \text{[illegible]} \bar{X}k_i)(\alpha + \beta X_i + u_i) \text{[illegible]}$$

$$\sum_{i=1}^{n} \frac{1}{n}\alpha + \beta\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} u_i - \alpha\bar{X}\sum_{i=1}^{n} k_i - \beta\bar{X}\sum_{i=1}^{n} k_i X_i - \bar{X}\sum_{i=1}^{n} k_i u_i$$

Since $\sum k_i = 0$, $\sum k_i X_i = 1$, $\sum 1 = n$ we have

$\hat{\alpha} = \alpha + \beta\bar{X} + \frac{1}{n}\sum_{i=1}^{n} u_i - \beta\bar{X} - \bar{X}\sum_{i=1}^{n} k_i u_i$

or $\hat{\alpha} = \alpha + \frac{1}{n}\sum_{i=1}^{n} u_i - \bar{X}\sum_{i=1}^{n} k_i u_i$ or $E(\hat{\alpha}) = E(\alpha) + \frac{1}{n}\sum_{i=1}^{n} E(u_i) - \bar{X}\sum_{i=1}^{n} k_i E(u_i)$

$E(\hat{\alpha}) = \alpha$ as $E(u_i) = 0$

This shows that mean of $\hat{\alpha}$ is α.

Thus it is proved that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and

1. The minimum variance property

In this property we shall prove the Gauss-Markov Theorem, which states that the least squares estimates are best (have the smallest variance) as compared with any other linear unbiased estimator obtained from other econometric methods.

First we have to find Var($\hat{\beta}$) and Var($\hat{\alpha}$) and then we have to prove the minimum variance property.

Variance of $\hat{\beta}$ = Var($\hat{\beta}$) = $E[\hat{\beta} - E(\hat{\beta})]^2 = E(\hat{\beta} - \beta)^2$ as $E(\hat{\beta})$ [illegible]

Since $\hat{\beta} = \beta + \sum_{i=1}^{n} K_i u_i$ (see property 2)

$$\hat{\beta} - \beta = \sum_{i=1}^{n} K_i u_i$$

or, $(\hat{\beta} - \beta)^2 = \left(\sum_{i=1}^{n} K_i u_i\right)^2$ or, $E(\hat{\beta} - \beta)^2 = E\left[\sum_{i=1}^{n} K_i u_i\right]^2$

or, $Var(\hat{\beta}) = E\left[\sum_{i=1}^{n} K_i^2 u_i^2 + 2\sum_{i<j} \sum K_i K_j u_i u_j\right]$

$$= \sum_{i=1}^{n} K_i^2 E(u_i^2) + 2\sum_{i<j} \sum K_i K_j E(u_i u_j)$$

$$= \sum_{i=1}^{n} K_i^2 E(u_i^2) \quad [\because E(u_i u_j) = 0, \text{ for } i \neq j]$$

$$= \sum_{i=1}^{n} K_i^2 \sigma_u^2 \quad [\because E(u_i^2) = \sigma_u^2]$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sigma_u^2 \quad \left[\because K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}\right]$$

$$= \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \quad \therefore Var(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

Similarly, Variance of $\hat{\alpha} = Var(\hat{\alpha})$

Since $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ (see property 1)

Substituting $\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$ we obtain $\hat{\alpha} = \bar{Y} - \bar{X}\sum_{i=1}^{n} K_i Y_i$

$$= \frac{\sum_{i=1}^{n} Y_i}{n} - \bar{X}\sum_{i=1}^{n} K_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}K_i\right) Y_i$$

Now $\quad \operatorname{var}(\hat{\alpha}) = \operatorname{var}\left[\sum\left(\frac{1}{n} - \bar{X}k_i\right)Y_i\right] = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)^2 \operatorname{var}(Y_i)$

Since $\operatorname{var}(Y_i) = \sigma_u^2$

$$\operatorname{Var}(\hat{\alpha}) = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)^2 \sigma_u^2$$

$$= \sigma_u^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)^2 = \sigma_u^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} - \frac{2}{n}\bar{X}k_i + \bar{X}^2k_i^2\right)$$

$$= \sigma_u^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right] \quad \left[\sum_{i=1}^{n} k_i = 0 \text{ and } \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2} \text{ where } k_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}\right]$$

$$= \sigma_u^2\left[\frac{\sum_{i=1}^{n} x_i^2 + n\bar{X}^2}{n\sum_{i=1}^{n} x_i^2}\right] = \sigma_u^2 \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}$$

$$\operatorname{Var}(\hat{\alpha}) = \sigma_u^2 \sum_{i=1}^{n} X_i^2 \Big/ n\sum_{i=1}^{n} x_i^2$$

$$\left[\sum_{i=1}^{n} x_i^2 + n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 + n\bar{X}^2\right.$$
$$\left. = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + 2n\bar{X}^2 = \sum_{i=1}^{n} X_i^2\right]$$

Case (a) $\hat{\beta}$ has the least variance.

We know that $\operatorname{Var}(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2$

Now we want to prove that any other linear unbiased estimate of the true parameter for example β^* obtained from any other econometric method, has a bigger variance than the least squares estimate $\hat{\beta}$. Thus we have to prove that $\operatorname{Var}(\hat{\beta}) < \operatorname{Var}(\beta^*)$.

Proof. The new estimator β^* is by assumption a linear combination of the Y_i's, a weighted sum of the sample values Y_i, the weights $k_i = x_i \Big/ \sum x_i^2$ being different from the weights of the least-squares estimates.

For example, let us assume $\beta^* = \sum_{i=1}^{n} C_i Y_i$ where $C_i = K_i + d_i$, d_i is an arbitrary set of weights similar to (but not the same as) to the K_i's.

Now put $Y_i = \alpha + \beta X_i + u_i$ in the expression of β^* and we obtain

$$\beta^* = \sum_{i=1}^{n} C_i(\alpha + \beta X_i + u_i) = \sum_{i=1}^{n} (\alpha C_i + \beta C_i X_i + C_i u_i)$$

It is assumed that the β^* is also an unbiased estimator of β i.e. $E(\beta^*) = \beta$

Now $E(\beta^*) = E\left[\sum_{i=1}^{n} (\alpha C_i + \beta C_i X_i + C_i u_i)\right]$

$$E(\beta^*) = E\left[\alpha \sum_{i=1}^{n} C_i + \beta \sum_{i=1}^{n} C_i X_i + \sum_{i=1}^{n} C_i u_i\right]$$

Now $E(\beta^*) = \beta$ if, and only if

$$\sum_{i=1}^{n} C_i = 0, \quad \sum_{i=1}^{n} C_i X_i = 1 \text{ and } \sum_{i=1}^{n} C_i u_i = 0$$

But $\sum_{i=1}^{n} C_i = 0$ implies $\sum_{i=1}^{n} d_i = 0$ because

$$\sum_{i=1}^{n} C_i = \sum_{i=1}^{n} (K_i - d_i) = \sum_{i=1}^{n} K_i + \sum_{i=1}^{n} d_i \text{ and } \sum_{i=1}^{n} K_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0 \text{ as } \sum_{i=1}^{n} x_i = 0$$

i.e., $\sum_{i=1}^{n} C_i = \sum_{i=1}^{n} d_i$. Therefore if $\sum_{i=1}^{n} C_i = 0$, then $\sum_{i=1}^{n} d_i = 0$

Similarly $\sum_{i=1}^{n} C_i X_i = 1$ requires $\sum_{i=1}^{n} d_i X_i = 0$,

since $\sum_{i=1}^{n} C_i X_i = \sum_{i=1}^{n} (K_i + d_i) X_i = \sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} d_i X_i$

Given that $\sum_{i=1}^{n} K_i X_i = 1$, $\sum_{i=1}^{n} C_i X_i = 1$ if $\sum_{i=1}^{n} d_i X_i = 0$

Thus β^* will be a linear unbiased estimate of β (with weights $C_i = K_i + d_i$) if

$$\sum_{i=1}^{n} C_i = 0, \quad \sum_{i=1}^{n} d_i = 0, \quad \sum_{i=1}^{n} C_i X_i = 1 \text{ and } \sum_{i=1}^{n} d_i X_i = 0$$

Now [illegible]

[illegible]

[illegible]

[illegible] where $Y_i = \alpha + \beta X_i + u_i$

$$E(Y_i) = E(\alpha + \beta X_i + u_i)$$

$$= \alpha + \beta X_i \quad \text{as } E(u_i) = 0$$

Now $var(Y_i) = E[Y_i - E(Y_i)]^2 = E[Y_i - \alpha - \beta X_i]^2$

$$= E(u_i^2) = \sigma_u^2$$

Similarly we may obtain

$$\beta^* = \sum_{i=1}^{n} c_i Y_i \text{ and } Var(\beta^*) = Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \, Var(Y_i) = \sum_{i=1}^{n} c_i^2 \sigma_u^2$$

$$\text{Now } \sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} (k_i + d_i)^2$$

$$= \sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} k_i d_i = \sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2$$

$$\text{Given that } \sum_{i=1}^{n} k_i d_i = \frac{\sum_{i=1}^{n} x_i d_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) d_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} d_i X_i - \bar{X}\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} x_i^2} = 0 \text{ as } \sum_{i=1}^{n} d_i X_i = 0 \text{ and } \sum_{i=1}^{n} d_i = 0$$

Substituting we find

$$Var(\beta^*) = \sigma_u^2 \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2\right) = \sigma_u^2 \sum_{i=1}^{n} k_i^2 + \sigma_u^2 \sum_{i=1}^{n} d_i^2$$

$$Var(\beta^*) = Var(\hat{\beta}) + \sigma_u^2 \sum_{i=1}^{n} d_i^2 \quad \left[Var(\hat{\beta}) = \sigma_u^2 \sum_{i=1}^{n} k_i^2 = \sigma_u^2 / \sum x_i^2 \right]$$

Since $\sigma_u^2\sum^{n}$ [illegible] it proves that $Var(\beta^*)$ [illegible] $Var(\hat\beta)$ or $Var(\hat\beta) < Var(\beta^*)$

Thus it is proved that $\hat\beta$ is the BLUE of β

Case (b) In the same way it can be proved that the least squares constant intercept $\hat\alpha$ possesses minimum variance. We take a new estimator α^* which we assume to be a linear function of the Y's with weights $C_i = k_i + d_i$

where $k_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$

Since $\hat\alpha = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}k_i\right)Y_i$

Similarly, $\alpha^* = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}C_i\right)Y_i$

This shows that like $\hat\alpha$, α^* is also a linear function in Y's
Now α^* is to be regarded as an unbiased estimator of α ($E(\alpha^*) = \alpha$)
We substitute for $Y_i = \alpha + \beta X_i + u_i$ in α^* and we get,

$$\alpha^* = \alpha\left[1 - \bar{X}\sum_{i=1}^{n} C_i\right] + \beta\left[\bar{X} - \bar{X}\sum_{i=1}^{n} C_iX_i\right] + \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}C_i\right)u_i$$

$$\text{Now } E(\alpha^*) = \alpha\left[1 - \bar{X}E\left(\sum_{i=1}^{n} C_i\right)\right] + \beta\left[\bar{X} - \bar{X}E\left(\sum_{i=1}^{n} C_iX_i\right)\right] + E\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}C_i\right)u_i\right]$$

Now $E(\alpha^*) = \alpha$ if and only if $\sum_{i=1}^{n} C_i = 0$, $\sum_{i=1}^{n} C_iX_i = 1$ and $\sum_{i=1}^{n} C_iu_i = 0$

These conditions imply $\sum_{i=1}^{n} d_i = 0$ and $\sum_{i=1}^{n} d_iX_i = 0$

The variance of α^* is given by
$Var(\alpha^*) = E[\alpha^* - E(\alpha^*)]^2 = E[\alpha^* - \alpha]^2$

$$= \sigma_u^2\sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}C_i\right]^2 = \sigma_u^2\sum\left[\frac{1}{n^2} - 2\frac{1}{n}\bar{X}C_i + \bar{X}^2C_i^2\right]$$

$$= \sigma_u^2\left[\frac{n}{n^2} - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n} C_i + \bar{X}^2\sum_{i=1}^{n} C_i^2\right] = \sigma_u^2\left[\frac{1}{n} + \bar{X}^2\sum^{n} C_i^2 - \frac{2}{n}\bar{X}\sum_{i=1}^{n} C_i\right]$$

Since $\sum_{i=1}^{n} C_i = 0$ and $\sum_{i=1}^{n} C_i^2 = \sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2$

we have $Var(\alpha^*) = \sigma_u^2 \left[\frac{1}{n} + \bar{X}^2 \sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right]$

$$= \sigma_u^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right] + \sigma_u^2 \left(\sum_{i=1}^{n} d_i^2 \right), \text{ where } \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n} x_i^2}$$

$$Var(\alpha^*) = \sigma_u^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right] + \sigma_u^2 \left(\sum_{i=1}^{n} d_i^2 \right) \text{ or, } Var(\alpha^*) = Var(\hat{\alpha}) + \sigma_u^2 \sum_{i=1}^{n} d_i^2$$

Here $\sum_{i=1}^{n} d_i^2 > 0$, because all d_i s are not zero

Thus we have, $Var(\alpha^*) > Var(\hat{\alpha})$ or, $Var(\hat{\alpha}) < Var(\alpha^*)$

Hence it is proved that $\hat{\alpha}$ is the BLUE of α.

2.10. The Variance of the Random Variable, u

The formulae of the variance of $\hat{\alpha}$ and $\hat{\beta}$ involve the variance of the random term u, σ_u^2. However, the true variance of u_i cannot be computed since the values of u_i are not observable. But we may obtain an unbiased estimate of σ_u^2 from the expression

$\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2)$ where $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$

{Y is the observed value and $\hat{Y}_i$ is the estimated value i.e. $Y = \alpha + \beta X_i + e_i$ and $\hat{Y} = \hat{\alpha} + \hat{\beta} X_i$ for $i = 1, 2, \ldots, n$}

Proof One property of the regression line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ is that it passes through the point $(\bar{X}, \bar{Y})$. So, $\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}$

Again we know that $\bar{Y} = \alpha + \beta \bar{X} + \bar{u}$ from the observed relationship

$\Bigg[$Where $Y_i = \alpha + \beta X_i + u_i$ $\quad \sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} u_i$

or, $\sum_{i=1}^{n} Y_i \Big/ n = \alpha + \beta \sum_{i=1}^{n} X_i \Big/ n + \sum_{i=1}^{n} u_i \Big/ n$ or, $\bar{Y} = \alpha + \beta \bar{X} + \bar{u} \Bigg]$

Since $e_i = Y_i - \hat{Y}_i$

$= (Y_i - \bar{Y}) + (\bar{Y} - \hat{Y}_i) = (\alpha + \beta X_i + u_i) - (\alpha + \beta\bar{X} + \bar{u}) - (\hat{\alpha} + \hat{\beta} X_i - \hat{\alpha} - \hat{\beta}\bar{X})$

$= [\beta(X_i - \bar{X}) + (u_i - \bar{u})] - [\hat{\beta}(X_i - \bar{X})]$

$e_i = -(\hat{\beta} - \beta)x_i + (u_i - \bar{u})$ where $x_i = X_i - \bar{X}$

or $e_i^2 = (\hat{\beta} - \beta)^2 x_i^2 + (u_i - \bar{u})^2 - 2x_i(\hat{\beta} - \beta)(u_i - \bar{u})$

$= (\hat{\beta} - \beta)^2 x_i^2 + (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta)(x_i u_i - x_i \bar{u})$

$$\sum_{i=1}^{n} e_i^2 = (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta)\left(\sum_{i=1}^{n} x_i u_i - \bar{u}\sum_{i=1}^{n} x_i\right)$$

$$= \left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sum_{i=1}^{n} x_i^2 + \left[\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i\right)^2}{n}\right] - 2\frac{\left(\sum_{i=1}^{n} x_i u_i\right)^2}{\sum_{i=1}^{n} x_i^2}$$

Since $\hat{\beta} = \beta + \dfrac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$, $\hat{\beta} - \beta = \sum_{i=1}^{n} x_i u_i \Big/ \sum_{i=1}^{n} x_i^2$

Again, $\sum_{i=1}^{n} (u_i - \bar{u})^2 = \sum_{i=1}^{n} u_i^2 - 2\bar{u}\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} \bar{u}^2$

$= \sum_{i=1}^{n} u_i^2 - 2\bar{u}\, n \frac{1}{n}\sum_{i=1}^{n} u_i + n\bar{u}^2 = \sum_{i=1}^{n} u_i^2 - 2n\bar{u}^2 + n\bar{u}^2$

$= \sum_{i=1}^{n} u_i^2 - n\bar{u}^2 = \sum_{i=1}^{n} u_i^2 - n\left(\sum_{i=1}^{n} \frac{u_i}{n}\right)^2 = \sum_{i=1}^{n} u_i^2 - \frac{\left[\sum u_i\right]^2}{n}$

and $2(\hat{\beta} - \beta)\left(\sum_{i=1}^{n} x_i u_i - \bar{u}\sum_{i=1}^{n} x_i\right) = 2\dfrac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i u_i = 2\left(\sum_{i=1}^{n} x_i u_i\right)^2 \Big/ \sum_{i=1}^{n} x_i^2$, as $\sum x_i = 0$

$$\sum_{i=1}^{n} e_i^2 = \left(\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i\right)^2}{n}\right) - \left(\sum_{i=1}^{n} x_i u_i\right)^2 \Big/ \sum_{i=1}^{n} x_i^2$$

or $\sum_{i=1}^{n} e_i^2 = \sum$ [illegible]

$E\left[\sum_{i=1}^{n} e_i^2\right] = \sum$ [illegible]

$= \left[\sum_{i=1}^{n} \sigma_u^2 - \sum \frac{\sigma_u^2}{n}\right.$ [illegible] $\left.\right]$ [as $E(u_i^2) = \sigma_u^2$ and $E(u_i u_j) = 0$ [illegible]]

or $E\left(\sum_{i=1}^{n} e_i^2\right) = n\sigma_u^2 - \frac{n\sigma_u^2}{n}$ [illegible]

$= n\sigma_u^2 - \sigma_u^2$ [illegible] $= \sigma_u^2(n-2)$

or, $E\left[\frac{\sum_{i=1}^{n} e_i^2}{n-2}\right] = \sigma_u^2$

So, $\sum_{i=1}^{n} e_i^2/(n-2)$ is an unbiased estimator of σ_u^2. If we denote $\sum_{i=1}^{n} e_i^2/(n-2) = \hat{\sigma}_u^2$, then $\hat{\sigma}_u^2$ is an unbiased estimator of σ_u^2.

2.8. Maximum Likelihood Estimators (MLE's) of α, β and σ_u^2

If each $u_i\,(Y_i = \alpha + \beta X_i + u_i)$ is normally distributed with mean 0 and variance σ_u^2 i.e., $u_i \sim N(0, \sigma_u^2)$ and $u_1, u_2, \ldots, u_n$ are independent, then MLE of α and β are equivalent to the OLS estimators of α and β (i.e., $\hat{\alpha}$ and $\hat{\beta}$).

Proof. Since $u_i \sim N(0, \sigma_u^2)$, the p.d.f. of u_i is given by

$$f_i(u_i) = \frac{1}{\sqrt{2\pi}\,\sigma_u} e^{-\frac{1}{2}\left(\frac{u_i - 0}{\sigma_u}\right)^2} = \frac{1}{\sqrt{2\pi}\,\sigma_u} e^{-\frac{1}{2\sigma_u^2}u_i^2} \quad \text{as } E(u_i) = 0.$$

The joint probability distribution function of $u_1, u_2, \ldots, u_n$ is given by $f(u_1, u_2, \ldots, u_n)$ and given the set of sample observations it is looked upon as a function of parameters and is called the [illegible] function of the parameters. Since $u_1, \ldots, u_n$ are independent, then we can write

[illegible]

[illegible]

[illegible]

Taking Log on both sides we get,

$$\log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}u_i^2$$

$$= -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$$

$$\left[\because Y_i = \alpha + \beta X_i + u_i \text{ or } u_i = Y_i - \alpha - \beta X_i \text{ or } \sum_{i=1}^{n}u_i^2 = \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2\right]$$

MLE of α and β can be obtained by maximising $\log L$ through the choice of α and β. Maximisation of $\log L$ through the choice of α and β is equivalent to minimisation of $\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$ through the choice of α and β.

Let us suppose that α^* is the MLE of α and β^* is the MLE of β. Then,

$$\beta^* = \frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2}, \quad \alpha^* = \bar{Y} - \beta^*\bar{X} \qquad \left[\text{Since } \hat{\beta} = \sum_{i=1}^{n}x_i y_i \Big/ \sum_{i=1}^{n}x_i^2 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}\right]$$

$$\text{Since } \log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$$

Differentiating partially $\log L$ with respect to α, β we get,

$$\frac{\delta \log L}{\delta\alpha} = -\frac{1}{2\sigma_u^2}\,2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)(-1)$$

$$\frac{\delta \log L}{\delta \log\hat{\beta}} = -\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)(-X_i)$$

Equating these equations to zero and putting star mark on the parameters to distinguish them from least squares estimators

we get

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}(Y_i-\alpha^*-\beta^*X_i)=0 \qquad (1)$$

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}X_i(Y_i-\alpha^*-\beta^*X_i)=0 \qquad (2)$$

The first two equations are reduced to the least squares normal equations

$$\sum_{i=1}^{n}Y_i=n\alpha^*+\beta^*\sum_{i=1}^{n}X_i$$

$$\sum_{i=1}^{n}X_iY_i=\alpha^*\sum_{i=1}^{n}X_i+\beta^*\sum_{i=1}^{n}X_i^2$$

Now solving the two normal equations we can get $\beta^*=\dfrac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}$ and $\alpha^*=\bar{Y}-\beta^*\bar{X}$.

This proves that the MLE of α and β are the same as the least squares estimators. Hence they would also possess all the desirable properties.

When log L is maximised through the choice of α and β and α* and β* are the MLEs of α and β, then,

$$\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}(Y_i-\alpha^*-\beta^*X_i)^2 \qquad [\because \alpha^*=\hat{\alpha} \text{ and } \beta^*=\hat{\beta}]$$

Hence, $$\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}X_i)^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}e_i^2 \qquad [\because \hat{Y}=\hat{\alpha}+\hat{\beta}X_i]$$

$$\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}e_i^2$$

So, the likelihood function maximised with respect to α and β is given by

$$\log L=-\frac{n}{2}\log 2\pi-n\log\sigma_u-\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}e_i^2$$

In order to obtain the MLE of σ_u^2, log L is to be maximised through the choice of σ_u and the first order condition of maximisation is given by (assuming σ_u^{2*} as the MLE of σ_u^2).

$$\frac{\delta\log L}{\delta\sigma_u}=-\frac{n}{\sigma_u}-\frac{1}{2}\sum_{i=1}^{n}e_i^2(-2)\frac{1}{\sigma_u^3}=0=\frac{1}{\sigma_u}\left[-n+\sum_{i=1}^{n}e_i^2\Big/\sigma_u^2\right]=0$$

or, $\sum_{i=1}^{n}e_i^2\Big/\sigma_u^2=n$ or, $\sigma_u^2=\sum_{i=1}^{n}e_i^2\Big/n=\sigma_u^{2}{}^*$ (say)

So $\sum e_i^2/n = \sigma_u^2*$ which is the MLE of the variance of the disturbance term denoted by σ_u^2*

Note: For maximisation, however, we require second order conditions which have not shown here. But we should also check the second order conditions required for maximisation. This means that we have to show that $\frac{\delta^2 \log L}{\delta \alpha^2} < 0$, $\frac{\delta^2 \log L}{\delta \beta^2} < 0$ and $\frac{\delta^2 \log L}{\delta \sigma_u^2} < 0$

Thus we see that the MLE of σ_u i.e. $\sigma_u^2* = \sum e_i^2/n$ is not an unbiased estimator but it is a consistent estimator.

$$\text{i.e. } E\left(\sum_{i=1}^{n} e_i^2/n\right) = E\left[\sum_{i=1}^{n} e_i^2/(n-2)\right]\cdot\frac{n-2}{n}$$

$$= E\left[\sum_{i=1}^{n} e_i^2/(n-2)\left(\frac{n-2}{n}\right)\right] = E\left[\sum_{i=1}^{n} e_i^2/(n-2)\right]\left(1-\frac{2}{n}\right)$$

$$= \sigma_u^2\left(1-\frac{2}{n}\right) \text{ since } E\left[\sum_{i=1}^{n} e_i^2/(n-2)\right] = \sigma_u^2$$

$$E\left(\sum_{i=1}^{n} e_i^2/n\right) = \sigma_u^2\left(1-\frac{2}{n}\right)$$

$$\text{Now } E\left[\sum_{i=1}^{n} e_i^2/n\right] \to \sigma_u^2 \text{ as } n \to \infty$$

This proves that the MLE of σ_u^2 i.e. $\sum_{i=1}^{n} e_i^2/n$ is a consistent estimator of σ_u^2

Note: MLE of α and β i.e., $\alpha*$ and $\beta*$ are unbiased estimators of α and β. But MLE of σ_u^2 i.e. $\sigma_u^2* = \sum_{i=1}^{n} e_i^2/n$ is not an unbiased estimator, rather it is a consistent estimator (consistently unbiased) of σ_u^2

2.12 The Sampling Distribution of the Least Squares Estimates

Since least squares estimators are linear combinations of independent normal variables, $Y_1, Y_2, \ldots, Y_n$, $\hat{\alpha}$ and $\hat{\beta}$ must also be normally distributed with the following characteristics [illegible] being equal to their true values of $\hat{\alpha}$ and $\hat{\beta}$.

[illegible]

[illegible]

$$\hat{\alpha} \sim N\left[\alpha, \sigma_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\right]; \quad E(\hat{\alpha}) = \alpha; \quad \text{var}(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right)$$

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}\right); \quad E(\hat{\beta}) = \beta; \quad \text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

Variances of the parameters are directly related to the variances of the disturbances. The following points should be noted carefully:

(i) Larger the value of σ_u^2, the larger the variances of $\hat{\alpha}$ and $\hat{\beta}$. In other words the greater the dispersion of the disturbance terms around the population regression line, the greater is the dispersion in the values of estimated regression parameters.

(ii) $\sum_{i=1}^{n} x_i^2$ is in the denominator of the variance formula of both the estimators. This indicates that the more dispersed the values of the explanatory variables, i.e., the greater $\sum_{i=1}^{n} x_i^2$, the smaller the variances of $\hat{\alpha}$ and $\hat{\beta}$. If $\sum_{i=1}^{n} x_i^2$ tends to zero, i.e., when $X_1 = X_2 = \ldots = X_n$, both variances would be infinitely large.

(iii) The variance of $\hat{\alpha}$ is the smallest when $\bar{X} = 0$ or tends to zero. In particular when $\bar{X} = 0$, $\text{var}(\hat{\alpha}) = \frac{\sigma_u^2}{n}$.

2.13 Confidence Intervals and Hypothesis Testing

It is highly essential to construct confidence intervals of the parameters in order to achieve precision of $\hat{\alpha}$ and $\hat{\beta}$. We have all the information concerning the distribution of $\hat{\alpha}$ and $\hat{\beta}$ in order to standardise them.

Since $\hat{\alpha} \sim N\left[\alpha, \sigma_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)\right]$ and $\hat{\beta} \sim N\left(\beta, \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}\right)$

$$t_{(n-2)} \text{ or } Z = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = \frac{\hat{\beta} - \beta}{\sigma_u \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} \sim N(0, 1)$$

where $SE(\hat{\beta}) = \sqrt{\operatorname{var}(\hat{\beta})}$ and t or $Z = \frac{\hat{\alpha} - E(\hat{\alpha})}{SE(\hat{\alpha})} = \frac{\hat{\alpha} - \alpha}{\sigma_u \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}}} \sim N(0, 1)$

Here $\operatorname{var}(\hat{\alpha}) = \sigma_u^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right]$

$$SE(\hat{\alpha}) = \sqrt{\operatorname{var}(\hat{\alpha})} = \sigma_u \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}} = \sigma_u \sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n \sum_{i=1}^{n} x_i^2}$$

$$\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i^2 + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2}$$

$$= \frac{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \right] + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2} = \frac{\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right] + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{n} X^2 - 2n\bar{X}^2 + 2n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2} = \left[\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right]$$

Here σ_u^2 represents the variance of the unobservable disturbances which is unknown. In particular if σ_u^2 is not known and we substitute by its unbiased estimator

$\sum_{i=1}^{n}$ [illegible] $E(\hat{\sigma}_u^2)$ [illegible] $\sum$ [illegible] $= \sigma_u^2$ then the standard normal variate Z

or Z follows t distribution with $(n-2)$ degrees of freedom.

In case of $\hat{\alpha}$: Z

$$Z = \frac{\hat{\alpha} - \alpha}{\sigma_u\sqrt{\dfrac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{n\sum_{i=1}^{n} x_i^2}}{\sigma_u\sqrt{\sum_{i=1}^{n} X_i^2}}$$

When σ_u is not known and it is replaced by $\hat{\sigma}_u$, $\hat{\sigma}_u = \sqrt{\sum_{i=1}^{n} e_i^2/(n-2)}$, the unbiased estimator of σ_u, then we have,

$$t = t_{n-2} = \frac{(\hat{\alpha} - \alpha)\sqrt{n\sum_{i=1}^{n} x_i^2}}{\hat{\sigma}_u\sqrt{\sum_{i=1}^{n} X_i^2}} \text{ with d.f} = (n-2)$$

Now by rearranging in terms of t expression we have

$$\hat{\alpha} - \alpha = t_{n-2}\hat{\sigma}_u\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}} \quad \text{or} \quad \alpha = \hat{\alpha} \pm t_{n-2}\hat{\sigma}_u\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}$$

Therefore 95% confidence limits for α are

$$\hat{\alpha} \pm t_{0.025,\,n-2}\,\hat{\sigma}_u\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n\sum_{i=1}^{n} x_i^2}$$

Similarly, 99% confidence limits for α are

$$\hat{\alpha} \pm t_{0.005,\,n-2}\,\hat{\sigma}_u\sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n\sum_{i=1}^{n} x_i^2}$$

[The values of $t_{0.025,\,n-2}$ and $t_{0.005,\,n-2}$ corresponding to $(n-2)$ d.f can be obtained from the table, given at the end of the book.]

In the same way, for testing β, we have

$$z = \frac{\hat{\beta} - \beta}{\dfrac{\sigma_u}{\sqrt{\sum x_i^2}}} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{i=1}^{n} x_i^2}}{\sigma_u}$$

when σ_u is not known then it is replaced by its unbiased estimator $\hat{\sigma}_u$, then we have

$$t = t_{n-2} = \frac{(\hat{\beta} - \beta)\sqrt{\sum_{i=1}^{n} x_i^2}}{\hat{\sigma}_u}$$ with d.f. = $n - 2$. Now rearranging we may get

$$\beta = \hat{\beta} - t_{n-2}\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}} \text{ or } \beta = \hat{\beta} \pm t_{n-2}\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Therefore 95% confidence limits for β would be

$$\hat{\beta} \pm t_{0.025, n-2}\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

and 99% confidence limits for β would be

$$\hat{\beta} \pm t_{0.005, n-2}\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Confidence interval for σ_u^2 :

Under the normality assumption, the variable $\chi^2 = (n-2)\dfrac{\hat{\sigma}_u^2}{\sigma_u^2}$ follows a χ^2 distribution with d.f = $(n-2)$.

Therefore, we can use χ^2_{n-2} to establish a confidence interval for σ_u^2

$$P\left[\chi^2_{1-\frac{\alpha}{2}} \le \chi^2_{n-2} \le \chi^2_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

or, $$P\left[\chi^2_{1-\frac{\alpha}{2}} \le (n-2)\frac{\hat{\sigma}_u^2}{\sigma_u^2} \le \chi^2_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

[illegible]

Usual test procedure [illegible]

As such the null hypothesis [illegible] we have to test it against the alternative hypothesis $H_1: \beta \neq 0$ [illegible] would be the appropriate test statistic under H_0 [illegible]

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sigma_u \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta}\sqrt{\sum_{i=1}^{n} x_i^2}}{\sigma_u}$$

which follows a t-distribution with d.f. = $(n-2)$.

Now at 5% level of significance the null hypothesis will be rejected for the given sample if $|t_{n-2}|$ (observed) $> t_{0.025,\,n-2}$ and will be accepted otherwise i.e. [illegible]. Similarly at 1% level of significance the null hypothesis will be rejected for the given sample if $|t_{n-2}|$ (observed) $> t_{0.005,\,n-2}$ and will be accepted otherwise i.e. [illegible]. The confidence limits for β (acceptance region in a two tailed test) at 5% and 1% levels of significance with $(n-2)$ degrees of freedom will be given by,

$$-t_{0.025,\,n-2}\, SE(\hat{\beta}) \le \beta \le +t_{0.025,\,n-2}\, SE(\hat{\beta})$$

and $$-t_{0.005,\,n-2}\, SE(\hat{\beta}) \le \beta \le +t_{0.005,\,n-2}\, SE(\hat{\beta})$$

where $SE(\hat{\beta}) = \dfrac{\sigma_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$ (σ_u is not known and replaced by $\hat{\sigma}_u$)

2.13.1. The Exact Level of Significance: The *p*-value

We know that the significance level α in a hypothesis testing problem is the probability of making a Type-I error, i.e. α is the probability of rejecting the null hypothesis [illegible] which [illegible] the probability [illegible] the appropriate [illegible] levels [illegible] the test statistic [illegible] the tolerance [illegible] is rejected and otherwise accepted. [illegible] limit for the observed test statistic gets.

Now we can ask a natural question: what is the smallest value of significance level at which the null hypothesis gets rejected? The answer is p-value or probability value associated with the observed data set.

Once the p-value is computed, one can make use of [illegible] by comparing the p-value with α.

Usually, if the p-value ≤ α then reject H_0, otherwise accept H_0.

To illustrate it, we consider an example where we have fitted the regression [illegible] $Y = \alpha + \beta X$ showing the impact of education on wages, given a sample size $n = 13$, where Y = wages, X = Education years of schooling (the original data is not given here).

The estimated regression results are given below

$$Y = \alpha + \beta X$$

$$\Rightarrow \quad \hat{Y} = -0.0144 + 0.7240X$$

$$SE \quad (0.9317) \quad (0.0700), \quad r^2 = 0.9065$$

Suppose we like to test the null hypothesis $H_0: \beta = 0.5$ against the alternative hypothesis $H_1: \beta \neq 0.5$. The appropriate test statistic under $H_0: \beta = 0.5$ would be

$$t_{n-2} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = \frac{0.7240 - 0.5}{0.0700} = 3.2$$

$$t_{n-2}\,(\text{observed}) = 3.2$$

Now on the basis of the given sample $H_0: \beta = 0.5$ will be rejected at $100\alpha\%$ level of significance when $(H_1: \beta \neq 0.5)$ if $|t_{n-2}\,(\text{observed})| > t_{\alpha/2,\,n-2}$ (Table).

Here $n = 13$, if $\alpha = 0.05$, $t_{\alpha/2,\,n-2} = t_{0.025,11} = 2.201$

and if $\alpha = 0.01$, $t_{\alpha/2,\,n-2} = t_{0.005,11} = 3.106$

So, $H_0: \beta = 0.5$ is rejected both at 5% and 1% levels of significance as t_{n-2} (observed) $= 3.2 > 2.201$ and 3.106.

Now on the basis of p-value the null hypothesis will be rejected for the given sample if $p \leq \alpha$ and will be accepted otherwise.

Here given the null hypothesis that the true coefficient of education $\beta = 0.5$, we obtain a t value of 3.2. Now what is the p-value of obtaining a t value of as much as or greater than 3.2?

From the t table given in Appendix Table IX we see that for 11 d.f. the probability of obtaining such a t value must be smaller than 0.005 (one tail) or 0.01 (two tail

We [illegible] statistical packages [illegible] find that the p-value of [illegible] 3.2] = 0.0000 that [illegible] greater is about 0.0000 [illegible]. Here α = 0.0 [illegible] the p-value of the observed [illegible] hence [illegible].

The [illegible] of significance of the t statistic is much smaller than [illegible] conventional and arbitrary fixed level of significance such as [illegible] per cent [illegible] computed and [illegible].

As a matter of fact, if we were to use the p-value just computed and reject the null hypothesis that the true coefficient of education is 0, the probability of committing a Type I error would be only about 1 in [illegible].

For a given sample size, as t increases the p-value decreases, and we can therefore reject the null hypothesis with increasing confidence.

What is the relationship of the p-value to the level of significance α? If we make the habit of fixing α equal to the p-value of a test statistic (e.g., the t statistic), then there is no conflict between the two values. To put it differently, it is better to give up fixing α arbitrarily at some level and simply choose the p-value of the test statistic. [illegible] we reject the null hypothesis, H_0, on the basis of the given sample if p-value [illegible] and accept H_0 otherwise.

2.14 Goodness of Fit of the Multiple Correlation Coefficient (R^2)

So far we were concerned with the estimation and precision of the regression parameters α and β. We now like to consider the regression line as a whole and examine its goodness of fit. Suppose, a sample regression has been obtained by the method of least squares, as shown in the following diagram (Fig. 2.5). Considering a specific observation of the dependent variable Y_i, we can write $e_i = Y_i - \hat{Y}_i$ where $Y_i = \alpha + \beta X_i + e_i$ and $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, e_i being the error of estimate.

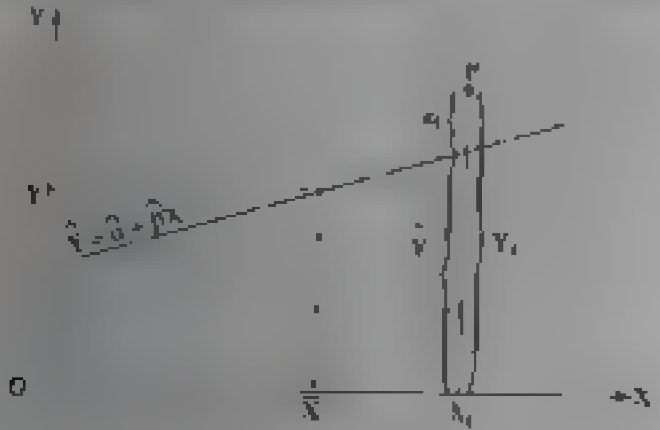


Fig. 2.5

Now $Y_i = \hat{Y}_i + e_i$

or $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i - \bar{e}$ or $(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + e_i$ ($\because \bar{e} = 0$)

or $\sum_{i=1}^{n} (Y_i - \bar{Y}) = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n} e_i$ and $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} [(\hat{Y}_i - \bar{Y}) + e_i]^2$

[illegible]

[illegible]

[illegible]

Now $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}X_i - \bar{Y})^2$

$$\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}X_i - \hat{\alpha} - \hat{\beta}\bar{X})^2 \quad [\because \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}]$$

$$\sum_{i=1}^{n}(\hat{\beta}X_i - \hat{\beta}\bar{X})^2 = \hat{\beta}^2\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i + \sum_{i=1}^{n}e_i^2$$

$$= \hat{\beta}^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + 2\times 0 + \sum_{i=1}^{n}e_i^2$$

or $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \hat{\beta}^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}e_i^2$

or $\sum_{i=1}^{n}y_i^2 = \hat{\beta}^2\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}e_i^2$ where $y_i = Y_i - \bar{Y}$ and $x_i = X_i - \bar{X}$

$$\Rightarrow \begin{bmatrix}\text{Total sum of}\\ \text{squares}\end{bmatrix} = \begin{bmatrix}\text{Explained sum}\\ \text{of squares}\end{bmatrix} + \begin{bmatrix}\text{Unexplained}\\ \text{sum of squares}\end{bmatrix} \text{or} \begin{bmatrix}\text{Residual sum}\\ \text{of squares}\end{bmatrix}$$

$\Rightarrow$ TSS = ESS + RSS

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ represents the total sum of squared deviations from $\bar{Y}$ which we may take as a measure of the total variations in Y.

Thus total variations can be decomposed into two parts

(i) $\hat{\beta}^2\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \hat{\beta}^2\sum_{i=1}^{n}x_i^2 \Rightarrow$ the estimated effect of change in X on the variations in Y (Explained sum of squares).

[illegible]

where R^2 [illegible]

$$[illegible] = 1 - \frac{RSS}{TSS}$$

Since $\text{var}(Y) = \text{var}(\hat{Y}) + \text{var}(e)$

and $0 \le \text{var}(\hat{Y}) \le \text{var}(Y)$

or $0 \le \frac{\text{var}(\hat{Y})}{\text{var}(Y)} \le 1$ or $0 \le R^2 \le 1$

$R^2 = 0$ when $\text{var}(\hat{Y}) = 0$ i.e. $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2$

and $R^2 = 1$ when $\text{var}(\hat{Y}) = \text{var}(Y)$ i.e. $\sum_{i=1}^{n} e_i^2 = 0$

It should be noted that, this R^2 is equal to the square of the simple correlation coefficient between X and Y.

By definition simple correlation coefficient (product moment) is given by

$$r_{XY} = r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Since $\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ and $R^2 = \frac{\beta^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$

$$R^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}\right)^2 = r^2$$

$$R^2 = r^2$$

Since $0 \le R^2 \le 1$

$$0 \le r^2 \le 1$$

or $-1 \le r \le +1$

Example 2.4. Find the value of R^2 from the following information and comment:

$\sum_{i=1}^{n} x_i y_i = 3347.60$, $\sum_{i=1}^{n} x_i^2 = 604.80$, $\sum_{i=1}^{n} y_i^2 = 19837$, $n = 30$, where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$

Solution Since $R^2 = \dfrac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$ where $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

$$\hat{\beta}^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)^2 = \left(\frac{3347.60}{604.80}\right)^2 = (5.54)^2 = 30.69$$

Now $$R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{30.69 \times 604.80}{19837} = \frac{18561.312}{19837} = 0.935$$

$$R^2 = 0.935$$

This suggests that 93.5 percent of the changes in the sample observations of Y can be attributed to the variations of the fitted value of Y i.e. $\hat{Y}$ or we say that our regression line fits the given data well.

Thus R^2 measures the proportion of variations in the dependent variable that is explained by the independent variables.

Example 2.5. A sample of 20 observations corresponding to the regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i is normally distributed with mean zero and unknown variance σ_u^2 gives the following data :

$$\sum_{i=1}^{n} Y_i = 21.9,\quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 86.9,\quad \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 106.4$$

$$\sum_{i=1}^{n} X_i = 186.2,\quad \sum_{i=1}^{n} (X_i - \bar{X})^2 = 215.4,\quad n = 20$$

Obtain the usual regression results.

Solution On the basis of the given information we have to fit a linear relation between Y (dependent variable) and X (explanatory variable).

i) **Estimation of α and β**

We know that, $$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta} = \frac{106.4}{215.4} = 0.494$$

and $\alpha = \bar{Y} - \beta\bar{X}$ where $\bar{Y} = \frac{\sum_{i=1}^{n} Y}{n} = \frac{21.9}{20} = 1.095$ and $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{186.2}{20} = 9.31$

$= 1.095 - 0.494 \times 9.31$

$= 1.095 - 4.60 = -3.505$

Thus we have, $\alpha = -3.505$ and $\beta = 0.494$ and our estimated regression line is

$\hat{Y} = \alpha + \beta X \Rightarrow \hat{Y} = -3.505 + 0.494X$

ii) **Estimation of variances**

Since we know that, $$\text{var}(\alpha) = \sigma_u^2\left(\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\right) \text{ and } \text{var}(\beta) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

Here we see that σ_u^2 is not known and hence we replace it by its unbiased estimator

$$\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / (n-2)$$

Thus we have, $$\text{var}(\alpha) = \hat{\sigma}_u^2\left(\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\right) \text{ and } \text{var}(\beta) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2}$$

Again we know that [illegible]

Since [illegible] where [illegible] and [illegible]

$$\sum e_i^2 = 86.9 - (0.494)^2 \times 215.4 = 86.9 - 52.56 = 34.34$$

Now $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / (n-2) = \frac{34.34}{20-2} = \frac{34.34}{18} = 1.908$

Now $\operatorname{var}(\hat{\alpha}) = \hat{\sigma}_u^2 \left(\frac{\sum X_i^2}{n \sum_{i=1}^{n} x_i^2} \right) = \frac{1.908 \times 1948.922}{20 \times 215.4} = 0.863$

$$\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 = 215.4 \text{ or } \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 215.4 \right.$$

or, $\sum_{i=1}^{n} X_i^2 = 215.4 + n\bar{X}^2 = 215.4 + 20 \times 9.31^2$

$= 215.4 + 1733.522 = 1948.922]$

$\therefore \operatorname{var}(\hat{\alpha}) = 0.8631$

Similarly, $\operatorname{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \frac{1.908}{215.4} = 0.0089$

Now, $SE(\hat{\alpha}) = \sqrt{\operatorname{var}(\hat{\alpha})} = \sqrt{0.8631} = 0.929$

$SE(\hat{\beta}) = \sqrt{\operatorname{var}(\hat{\beta})} = \sqrt{0.0089} = 0.094$

(iii) Construction of confidence intervals

Now we like to set up a confidence interval for α and β at (a) P = 0.95 (i.e., 5% level of significance) and (b) P = 0.99 (i.e., 1% level of significance).

In other words, we like to find the value of 't' that cuts off (a) 0.025 and (b) 0.005 of the area at the tail end of the distribution on both sides. From table value $t_{0.025}(n-2) = t_{0.025}(18) = 2.101$ and $t_{0.005}(n-2) = t_{0.005}(18) = 2.878$.

Therefore 95% confidence interval for α are $\hat{\alpha} \pm t_{0.025}(n-2)\, SE(\hat{\alpha})$ i.e., $P[\hat{\alpha} - t_{0.025}(n-2)\, SE(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.025}(n-2)\, SE(\hat{\alpha})] = 0.95$ and 99% confidence interval for α would be $\hat{\alpha} \pm t_{0.005}(n-2)\, SE(\hat{\alpha})$.

[illegible]

[illegible]

[illegible]

Thus 95% confidence interval for β would be [illegible]

or [illegible] .094 | where $\hat{\beta}$ = 0.494

or [illegible] | [illegible]

SE $(\hat{\beta})$ = 0.094

Hypothesis testing. Suppose we like to test the null hypothesis H_0 [illegible] against the alternative hypothesis $H_1: \beta \neq 0$. Now on the basis of the given data $H_0: \beta = 0$ will be rejected at 5% level of significance if

$$t_n = \frac{\hat{\beta}}{SE(\hat{\beta})} \text{ (observed)} > t_{0.025}\,(n-2) \text{ (table value)}$$

and may be accepted otherwise.

$$\text{Here } t_n = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.494}{0.094} = 5.255 \text{ (where } n = 20)$$

Thus we see that t_{n-2} = 5.255 > $t_{0.025}$ (18) = 2.101 and hence $H_0: \beta = 0$ is rejected (alternative $H_1: \beta \neq 0$ is accepted) at 5% level of significance. So the hypothesis of no relationship between X and Y is to be rejected at 5% level of significance. Similarly it can be tested for 1% level of significance.

2.6 Results of Regression Analysis

The results of regression analysis are generally presented in a conventional format. It is not sufficient merely to report the estimates of α and β. In practice we report regression coefficients together with their standard errors and the value of R^2. It has become customary to present the estimated equation with standard errors placed in parentheses below the estimated parameter values. These results are supplemented by R^2, the value of which is written on right hand side of the estimated regression equation.

In terms of our earlier example (Example 2.5) the estimated regression results can be written as

$$\hat{Y} = 3.505 + 0.494\,X \qquad R^2 = 0.6048$$

SE (0.929) (0.094)

Here [illegible]

[illegible]

This suggests that [illegible] per cent [illegible] be attributed to the variations of the [illegible] [illegible] the given data moderately [illegible]

[illegible] econometricians [illegible] report the t-ratios of the estimated coefficients instead of standard errors. This way of presentation makes the testing of hypothesis [illegible] direct.

Thus the other form of presentation of results [illegible]

$Y = 3.505 + 0.494X$ $R^2 = 0.668$

t-ratios (3.772) (5.255)

where $\alpha = 3.505$, $\beta = 0.494$

$$\frac{\alpha}{SE(\alpha)} = 3.772, \quad \frac{\beta}{SE(\beta)} = 5.255$$

Example 2.6. Suppose that Mr. X estimates a consumption function and obtains the results

$C = 15 + 0.81\, Y_d$ $n = 19$

t ratios (3.1) (18.7) $R^2 = 0.99$

C is consumption, Y_d is disposable income, the numbers in parentheses are t-ratios.

(a) Test the significance of Y_d statistically using t-ratios.

(b) Determine the estimated standard derivations of the parameter estimators.

(c) Construct a 95 percent confidence interval for the coefficient of Y_d.

Solution : It is a formal consumption function of the Keynesian type $C = a + bY_d$ where a = autonomous part of consumption and b = Marginal propensity to consume. By assumption in the existing theory $a > 0$, $0 < b < 1$. The estimated relation/regression results are given here as

$\hat{C} = 15 + 0.81\, Y_d$ $n = 19$

t ratios (3.1) (18.7) $R^2 = 0.99$

This shows that $\hat{a} = 15$, $\hat{b} = 0.81$,

$$\frac{\hat{a}}{SE(\hat{a})} = 3.1 \text{ (t-ratio) and } \frac{\hat{b}}{SE(\hat{b})} = 18.7 \text{ (t-ratio)}$$

n = no. of data points (sample size = 19)

R^2 = Square of multiple correlation coefficient

$$= \frac{ESS}{TSS} = \frac{\text{Explained variation in } C}{\text{Total variation}} = \frac{\text{var}(\hat{C})}{\text{var}(C)}$$

[illegible]

Here R^2 [illegible]

This means that [illegible] percent of the variations in sample observations of [illegible] are attributed to the variations of the fitted value of [illegible]. Thus we say that regression line fits the given data well. That of 100% variation in consumption [illegible] regression relation can explain [illegible]% variation in consumption.

(a) We have to test the null hypothesis $H_0: b = 0$ (no relation between C and [illegible]) against the alternative hypothesis $H_1: b \neq 0$.

The appropriate test statistic under $H_0: b = 0$ would be

$\frac{\hat{b}}{SE(\hat{b})}$ which follows a t distribution with $(n-2)$ degrees of freedom

i.e. $t_0 = t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ (given)

Now at 5% level of significance $H_0: b = 0$ (no relation between C and Y_d)

will be accepted if $-t_{0.025,\, n-2} \le t \le t_{0.025,\, n-2}$

and will be rejected otherwise.

From table value we get

$t_{0.025,\, n-2} = t_{0.025,\, 17} = 2.110$

($n = 19$ given)

Thus we see that observed $t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ does not lie in the interval -2.110 and 2.110 and hence the null hypothesis is rejected and the alternative is accepted. This means that there exists a relation between consumption (C) and disposable income [illegible]. Hence the relation is statistically significant.

(b) We have to find $SE(a)$ and $SE(b)$

Since for a, $t = \frac{\hat{a}}{SE(\hat{a})} = 3.1$ (given)

and $\hat{a} = 15$, $\quad 3.1 = \frac{15}{SE(\hat{a})}$ or, $SE(\hat{a}) = \frac{15}{3.1} = 4.8387$

Similarly, for b, $t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ (given)

and $\hat{b} = 0.81$

or, $18.7 = \frac{0.81}{SE(\hat{b})}$ or $SE(\hat{b}) = \frac{0.81}{18.7} = 0.0433$

Thus the estimated standard deviations of the parameter estimators are [illegible]

4. [illegible]

We have to construct 95 per cent confidence interval for the coefficient of [illegible] $(\alpha + \beta)$

This will be given by

[illegible]

[illegible]

[illegible]

[illegible] between [illegible] and [illegible]

This means that the coefficient of X_t will be between 0.7187 and 0.9013

Example 3.3 The following table shows data on labour hours of work and output for ten workers

X (labour hours of work)	10	7	10	5	8	8	6	7	9	10
Y (output)	11	10	12	6	10	7	9	10	11	10

(i) Assuming a linear regression of the form $Y_i = \alpha + \beta X_i + u_i$ [illegible] find the OLS estimators of α and β (i.e. $\hat{\alpha}$ and $\hat{\beta}$) and the estimated regression line $\hat{Y} = \hat{\alpha} + \hat{\beta}X$

(ii) Find $\text{var}(\hat{\alpha})$, $\text{var}(\hat{\beta})$, $SE(\hat{\alpha})$ and $SE(\hat{\beta})$

(iii) Find the value of $\sum_{i=1}^{n} e_i^2$ and $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / (n-2)$

(iv) Find the value of R^2

(v) Construct 95% confidence intervals of α, β and σ_u^2

(vi) Test the null hypothesis H_0: $\beta = 1.35$ against
(a) H_1: $\beta = 1.35$, (b) H_1: $\beta > 1.35$ (c) H_1: $\beta < 1.35$

Solution : **Calculations for the Regression**

Observations	X_i	Y_i	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x_i^2	$x_i y_i$	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
1	10	11	2	1.4	4	2.8	[illegible]	[illegible]
2	7	10	1	0.4	1	0.4	8.85	1.15
3	10	12	2	2.4	4	4.8	11.1	0.9
4	5	6	3	3.6	9	10.8	7.35	1.35
5	8	10	0	0.4	0	0	9.60	0.4
6	8	7	0	2.6	0	0	9.60	2.6
7	6	9	2	0.6	4	1.2	8.10	0.9
8	7	10	1	0.4	1	0.4	8.85	1.15
9	9	11	1	1.4	1	1.4	10.35	0.65
10	10	10	2	0.4	4	0.8	[illegible]	
Total	$\sum_{i=1}^{n} X_i = 80$	$\sum_{i=1}^{n} Y_i = 96$	$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 0$	$\sum_{i=1}^{n} x_i^2 = 28$	$\sum_{i=1}^{n} x_i y_i = 21$	$\sum_{i=1}^{n} \hat{Y}_i = 96$	$\sum_{i=1}^{n} e_i = 0$

Here $n = $ [illegible] $\bar{X} = \sum_{i=1}^{n} X_i / n = \frac{80}{10} = 8$, $\bar{Y} = \sum_{i=1}^{n} Y_i / n = \frac{96}{10} = 9.6$

$\sum_{i=1}^{n} e_i^2 = $ [illegible] $= 14.65$

and $\sum_{i=1}^{n} X_i^2 = $ [illegible] $= 668$

∴ the OLS estimators of α and β are given by $\hat{\alpha}$ and $\hat{\beta}$ where

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{21}{28} = 0.75$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 9.6 - 0.75 \times 8 = 9.6 - 6 = 3.6$

$\hat{\alpha} = 3.6$ and $\hat{\beta} = 0.75$

The estimated regression line equation becomes

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $\hat{Y} = 3.6 + 0.75X$. This equation is now used to find the values $\hat{Y}$ corresponding to different values of X. The values of $\hat{Y}$ are given in the above table showing calculations for regression.

Here the regression coefficient $\hat{\beta} = 0.75$ measures the marginal productivity of labour. The intercept $\hat{\alpha} = 3.6$ means that output will be 3.6 units when labour hours of work is zero.

(ii) We have to find $\text{var}(\hat{\alpha})$, $\text{var}(\hat{\beta})$, $SE(\hat{\alpha})$ and $SE(\hat{\beta})$.

Since we know that $\text{var}(\hat{\alpha}) = \sigma_u^2 \left(\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right)$ and $\text{var}(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2$

Since σ_u^2 is unknown it is replaced by its unbiased estimator

$$\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n - 2) = \frac{14.65}{10 - 2} = \frac{14.65}{8} = 1.8312$$

$$\text{Now } \text{var}(\hat{\alpha}) = \frac{\hat{\sigma}_u^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} = \frac{1.8312 \times 668}{10 \times 28} = \frac{1223.2416}{280} = 4.3687$$

and $SE(\hat{\alpha}) = \sqrt{\text{var}(\hat{\alpha})} = \sqrt{4.3687}$ [illegible]

[illegible]

[illegible]

[illegible] taken from last column of calculation table]

[illegible]

(iv) We have to find out the value of the [illegible] R

Since we know that $R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$ [Since $\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$

i.e., $TSS = ESS + RSS$]

$$\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$$

$$= (0.75)^2 \times 28 + 14.65 = 15.75 + 14.65 = 30.40$$

$$R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{(0.75)^2 \times 28}{30.40} = \frac{15.75}{30.40} = 0.51$$

This suggests that 51 percent of the variations in the sample observations of Y can be attributed to the variations of the fitted value of Y i.e. $\hat{Y}$. Here we see that our regression line fits the given data moderately (not very well).

From the above results we can write our regression results as follows

$$\hat{Y} = 3.6 + 0.75\,X,\ R^2 = 0.51$$

(2.090) (0.256) [SE values in brackets]

Alternatively, $\hat{Y} = 3.6 + 0.75\,X,\ R^2 = 0.51$

(1.7224) (2.930) [t values in brackets

(v) 95% confidence intervals of α, β and σ_u^2

(a) 95% confidence interval for α is given by,

$$P[\hat{\alpha} - t_{0.025,\,n-2}\,SE(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.025,\,n-2}\,SE(\hat{\alpha})] = 0.95$$

So 95% confidence limits of α are

$\hat{\alpha} \pm t_{0.025,\, n-2}\, SE(\hat{\alpha})$ [illegible]

Since $t_{0.025,\, n-2} = t_{0.025,\, 8} = 2.306$ from table value,

or [illegible] 4.8 [illegible]

and 8.4 [illegible]

So 95% confidence limits for α are [illegible] and 8.42

(b) 95% confidence interval for β is given by

$$P[\hat{\beta} - t_{0.025,\, n-2}\, SE(\hat{\beta}) < \beta < \hat{\beta} + t_{0.025,\, n-2}\, SE(\hat{\beta})] = 0.95$$

So 95% confidence limits of β are

$\hat{\beta} \pm t_{0.025,\, n-2}\, SE(\hat{\beta})$

or $0.75 \pm t_{0.025,\, 8}\, SE(\hat{\beta})$

or $0.75 \pm 2.306 \times 0.256$

or 0.75 ± 0.59

$\Rightarrow 0.16$ and 1.34

Since $t_{0.025,\, n-2} = t_{0.025,\, 8} = 2.306$ (From table value)

95% confidence limits of β are 0.16 and 1.34

(c) Since we know that $100(1-\alpha)$% confidence interval for σ_u^2 is given by

$$P\left[(n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,\, n-2}} \le \sigma_u^2 \le (n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,\, n-2}}\right] = 1-\alpha$$

Here $n = 10$, $\alpha = 0.05$, $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 1.8312$, $\chi^2_{\alpha/2,\, n-2} = \chi^2_{0.025,\, 8} = 17.535$ (Table value)

and $\chi^2_{1-\alpha/2,\, n-2} = \chi^2_{0.975,\, 8} = 2.180$ (Table value)

So, 95% confidence interval for σ_u^2 would be

$$P\left[8 \times \frac{(1.8312)}{17.535} \le \sigma_u^2 \le 8 \times \frac{(1.8312)}{2.180}\right] = 0.95$$

or $P[0.84 < \sigma_u^2 < 6.72] = 0.95$

95% confidence limits of σ_u^2 are 0.84 and 6.72

(iv) To test the null hypothesis $H_0: \beta = 1.35$ against the alternative hypothesis $H_1: \beta \ne 1.35$ the appropriate test statistic would be, $t_{n-2} = \dfrac{\hat{\beta} - \beta}{SE(\hat{\beta})}$

Now on the basis of the sample data $H_0: \beta = 1.35$ will be rejected at 5% level of significance if

$$|t_{n-2}| = \left|\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}\right| \text{(observed)} > t_{0.025,\, n-2} \text{ (Table value)}$$

and will be accepted otherwise

Here $t_{n-2} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = \frac{0.75 - 1.35}{[illegible]} = \frac{-0.6}{[illegible]} = [illegible]$

$= -2.343$ and $t_{[illegible]} = 2.343$

From the table value we get [illegible]

Thus [illegible] 2.343 [illegible] This means that on the basis of sample [illegible] $H_0: \beta = 1.35$ is rejected at 5% level of significance.

(b) The null hypothesis $H_0: \beta = 1.35$ will be rejected against the alternative $H_1: \beta > 1.35$ at 100α% level of significance if for the given sample

$$t_{n-2} \text{ (observed)} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} > t_{\alpha, n-2} \text{ (table value)}$$

and will be accepted otherwise. Here $\alpha = 0.05$, $n = 10$

and t_{n-2} (observed) $= \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = -2.343 < t_{0.05, 8} = 1.860$

So $H_0: \beta = 1.35$ is accepted (hence insignificant) at 5% level of significance.

(c) The null hypothesis, $H_0 = 1.35$ will be rejected against the alternative $H_1: \beta < 1.35$ for the given sample at 100α% level of significance if

$$t_{n-2} \text{ (observed)} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} < -t_{\alpha, n-2} \text{ (table value)}$$

and will be accepted otherwise.

Here, t_{n-2} (observed) $= \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = \frac{0.75 - 1.35}{0.256} = -2.343$

$< t_{0.05, 8} = -1.860$ Here $\alpha = 0.05$, $n = 10$

This clearly shows that the null hypothesis $H_0: \beta = 1.35$ is rejected (significant) at 5% level of significance.

2.16. Analysis of Variance for the Simple Linear Regression Model

Yet another item that is often presented in connection with the simple linear regression model is the **analysis of variance.** This is the breakdown of the total sum of squares (TSS) into explained sum of squares (ESS) and the residual sum of squares (RSS). The purpose of presenting the table is to test the significance of the explained sum of squares. In this case this amounts to testing the significance of β.

In regression analysis, we minimise the square deviations from mean and it has been proved (see Section 2.14) that

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{ or } \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

That is, Total variation = Explained variation + Unexplained variation or Residual variance

or, TSS = ESS + RSS with degrees of freedom $(n - 1) = (K - 1) + (n - K)$ where

n = total number of observations (given) and

K = number of parameters to be estimated.

[illegible]

[illegible]

Now [illegible]

[illegible]

$$[illegible]$$

[illegible]

$$[illegible]$$

and hence we have

$$[illegible]$$

i.e. TSS = ESS + RSS

with d.f. [illegible] As here K = 2 as there are two parameters α & β.

Thus we see that total variations are split into explained by explanatory variable and unexplained error (residual) variations [illegible] between and within variations [illegible] analysis of variance procedure. This suggests that we can compile the analysis of variance type of table for the regression analysis also in order to [illegible] the overall significance of the regression results.

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares		
				Observed	Tabulated
Explained between	ESS $\hat{\beta}^2 \sum x_i^2$	K	ESS/(K − 1) = MSE	F = $\frac{MSE}{MSR}$ with d.f. = (K − 1, n − K)	
Residual within	RSS $\sum e_i^2$	n − K	RSS/(n − K) MSR		
Total	TSS = $\sum y_i^2$	n			

Here K = 2 as the model is a two variable regression model and two parameters are involved.

To test the null hypothesis concerning the [illegible] where the explanatory variable is single [illegible] we [illegible] F-test by computing F-ratio through the [illegible] sum of squares of explained variations to that [illegible] in testing H_0: β [illegible] against the alternative [illegible] we may use the test statistic,

$$F^* = \frac{MSE}{MSR} = \frac{\hat{\beta}^2 \sum x_i^2/(K-1)}{\sum e_i^2/(n-K)}$$

$$= \frac{\hat{\beta}^2 \sum x_i^2/1}{\sum e_i^2/(n-2)} \text{ with d.f. } 1, (n-2) \text{ (Here } K=2)$$

Now we have to compare F^*_{cal} with the tabulated value of F with d.f. 1 and $(n-2)$. If it is found that $F^* > F_{\alpha}$ (table), we reject the null hypothesis at [illegible] significance (α = 0.01 or 0.05) and accept otherwise. The relation will be significant when H_0 is rejected.

It should be noted that

$$F^* = \frac{MSE}{MSR} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum e_i^2/(n-2)} = \frac{(n-2)\hat{\beta}^2 \sum x_i^2}{(1-R^2)\sum y_i^2}$$

$$\text{Now } \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{\sum y_i^2 - \sum e_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

$$\text{But } \frac{\sum e_i^2}{\sum y_i^2} = 1 - R^2 \text{ or, } \sum e_i^2 = (1-R^2)\sum y_i^2 \Big]$$

$$\text{Therefore, } F^* = \frac{(n-2)\left\{\frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2}\right\}}{1-R^2} = \frac{(n-2)R^2}{1-R^2}$$

which, on generalisation becomes $\frac{R^2/(K-1)}{(1-R^2)/(n-K)}$ where we have K parameters.

Furthermore, we know that

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sqrt{\mathrm{var}(\hat{\beta})}} \quad \text{But } \mathrm{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2} = \frac{\sum e_i^2/(n-2)}{\sum x_i^2}$$

$$t^2 = \frac{\hat{\beta}^2}{\mathrm{var}(\hat{\beta})} = \frac{\hat{\beta}^2}{\left\{\sum e_i^2/(n-2)\right\}\left(\frac{1}{\sum x_i^2}\right)} \text{ or, } t^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum e_i^2/(n-2)} = F^*$$

The [illegible] and F tests are formally equivalent, the relation between [illegible] be [illegible]

Example 2.8. Let us consider the following data to construct the analysis of variance table for a simple regression model $Y_i = \alpha + \beta X_i + u_i$

Given $\sum_{i=1}^{n} Y_i = 21.9$, $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 86.9$, $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 106.4$

$\sum_{i=1}^{n} X_i = 186.2$, $\sum_{i=1}^{n} (X_i - \bar{X})^2 = 215.4$, $n = 20$

Solution (See Example 2.3)

The OLS estimators of α and β can be obtained as follows

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{106.4}{215.4} = 0.494$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$= 1.095 - 0.494 \times 9.31 = 1.095 - 4.60 = -3.505$

where $\bar{Y} = \sum_{i=1}^{n} Y_i / n = \frac{21.9}{20} = 1.095$ and $\bar{X} = \left(\sum_{i=1}^{n} X_i\right) / n = \frac{186.2}{20} = 9.31$

The estimated regression results are

$\hat{Y} = -3.505 + 0.494\,X$ $R^2 = 0.6048$

t ratios (-3.772) (5.255) where $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$

$SE(\hat{\alpha}) = 0.929$ $SE(\hat{\beta}) = 0.094$ $\frac{\hat{\alpha}}{SE(\hat{\alpha})} = -3.772$ $\frac{\hat{\beta}}{SE(\hat{\beta})} = 5.255$

$$R^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{(0.494)^2 \times 215.4}{86.9} = 0.6048$$

Now $\sum e_i^2 = \sum y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 = 86.9 - (0.494)^2 \times 215.4 = 34.34$

Now for testing $H_0: \beta = 0$

against the alternative $H_1: \beta \neq 0$, we may use the ANOVA Table

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-statistic	F-table value at 1% and 5%
Explained (by regression)	$ESS = \hat{\beta} \sum_{i}^{n} x_i y_i$ $= 52.56$	$k - 1$ $= 2 - 1 = 1$	$MSE = \frac{ESS}{k-1}$ $= \frac{52.56}{1}$	$F^* = \frac{MSE}{MSR}$ $= \frac{52.56}{1.908}$ $= 27.55$ with df (1, 18)	$F_{0.01}(1, 18)$ $= 8.29$
Residual (within)	$RSS = \sum_{i}^{n} e_i^2$ $= 34.34$	$n - k$ $= 20 - 2 = 18$	$MSR = \frac{RSS}{n-k}$ $= \frac{34.34}{18}$ $= 1.908$		$F_{0.05}(1, 18)$ $= 4.41$
Total	$TSS = \sum_{i=1}^{n} y_i^2$ $= 86.9$	$n - 1$ $= 20 - 1 = 19$			

Here we see that the observed $F^* = 27.55$ is much larger than table $F_{0.01}(1, 18) = 8.29$ and $F_{0.05}(1, 18) = 4.41$. This means that $H_0: \beta = 0$ is rejected both at 5% and 1% levels of significance. Hence, we reject the null hypothesis and accept that the regression is significant, that is, X is a significant explanatory factor of the variation in Y.

2.17 Testing the Equality between Coefficients Obtained from Different Regressions or Different Samples

Sometimes we may have to estimate a regression equation separately for several sets of data and we may have to test whether some or all the parameters are the same for all different sets of data.

Suppose, we have two samples on the variables Y and X containing n_1 observations for first set and, Y and X containing n_2 observations for second set. We may obtain two estimates of the same relationship for these two samples

$$\hat{Y}_1 = \hat{\alpha}_1 + \hat{\beta}_1 X$$

and $$\hat{Y}_2 = \hat{\alpha}_2 + \hat{\beta}_2 X$$

Now our problem is to examine whether these two estimated relations differ significantly. If yes, then we may conclude that the relationship changes from one sample to the other.

For example, suppose that we have the data on consumption and disposable income for the two periods 1990–1999 and 2000–2009. We estimate the consumption functions separately for these periods. Then we may be interested to examine whether the functions are statistically significant or whether the MPCs significantly differ or not.

[illegible]

Step 1 We have [illegible] the pooled [illegible] with number of observations [illegible] $\sum$ [illegible] and $\sum$ [illegible] $\sum$ [illegible] with [illegible] degrees of freedom [illegible] pooled data.

Step 2 Next we [illegible] the regression for each sample separately.

For first sample [illegible] and $\sum$ [illegible] $= \sum$ [illegible] $\sum$ [illegible]

For second sample [illegible] and $\sum$ [illegible] $= \sum$ [illegible] $\sum$ [illegible]

Step 3 Next we have to compute F^* value as follows

$F^* =$ [illegible] with d.f. $= k, (n_1 + n_2 - 2k)$

Next we have to test the null hypothesis

$H_0: \alpha_1 = \beta_1 =$ [illegible] against the alternative H_1: H_0 is not correct

[illegible] we reject the null hypothesis at 5% level of significance.

In particular, if the pooled results are not given then F^* which can be obtained as follows

$F^* =$ [illegible] with d.f. [illegible]

Example 2.9 In order to test the null hypothesis that there is no difference in the MPC (Marginal propensity to consume) of manual workers and white collar employees, a research team estimated the following consumption functions:

Manual workers Sample size $n_1 = 35$

$C_1 = 26 + 0.90Y$, $r^2 = 0.92$, $\sum (C_1 - \hat{C}_1)^2 = 3,251$

(2) (5.6)

(The numbers in brackets are the t values for the regression coefficients.)

White-colour employees Sample size $n_2 = 30$

$C_2 = 160 + 0.82Y$, $r^2 = 0.95$, $\sum (C_2 - \hat{C}_2)^2 = 4,532$

(2.3) (8.5)

The numbers in brackets are the t values for the regression coefficients.

Combined sample consumption function

sample size $n = n_1 + n_2 = 30 + 35 = 65$

$\hat{C} = 250 + 0.70Y$, $r^2 = 0.90$, $\sum e_p^2 = 6,330$

(5.3) (6.2)

On the basis of the above results, can we accept the hypothesis that there is no differences between the MPC's of the two groups (use a 5 per cent level significance)?

Solution Suppose the estimated consumption function for the first sample is [illegible] and for the second sample [illegible]

and the pooled estimated consumption function is [illegible]. We have to test the [illegible] hypothesis

$H_0: \beta_1 = \beta_2$ against the alternative H_1: H_0 is not true

where β_1 = MPC of the first sample

β_2 = MPC of the second sample

α_1 = MPC of the pooled sample)

The appropriate test statistic will be

$$F^* = \frac{\left[\sum e_p^2 - \left(\sum e_1^2 + \sum e_2^2\right)\right] / K}{\left(\sum e_1^2 + \sum e_2^2\right) / (n_1 + n_2 - 2K)} \text{ with}$$

$d.f. = [K, (n_1 + n_2 - 2K)]$

Now H_0 will be rejected (significant) if $F^* > F_{0.05}[K, (n_1 + n_2 - 2K)]$ and will be accepted otherwise.

Now on the basis of our given information we see that

From first sample

$n_1 = 35,\ r_1^2 = R_1^2 = 0.92,\ \sum (C_1 - \bar{C}_1)^2 = \sum c_1^2 = 3{,}251$

$K = 2,\ \hat{\beta}_1 = 0.90 \quad \sum e_1^2 = (1 - R_1^2)\sum c_1^2 = (1 - 0.92) \times 3{,}251 = 260.08$

$$\left[\text{since } R^2 = 1 - \frac{\sum e_i^2}{\sum c_i^2} \quad \frac{\sum e_i^2}{\sum c_i^2} = 1 - R^2 \text{ or } \sum e_i^2 = (1 - R^2)\sum c_i^2\right]$$

From second sample $n_2 = 30,\ \hat{\beta}_2 = 0.82 \quad r_2^2 = R_2^2 = 0.95$

$$\sum (C_2 - \bar{C}_2)^2 = \sum c_2^2 = 4{,}532,\ K = 2$$

and $\quad \sum e_2^2 = (1 - R_2^2)\sum c_2^2 = (1 - 0.95) \times 4{,}532 = 226.6$

From the pooled sample $n = n_1 + n_2 = 35 + 30 = 65$

$\hat{\alpha}_1 = 0.70,\ r^2 = R^2 = 0.92 \quad \sum e_p^2 = 16{,}320$

Thus we have, $\dfrac{\left[\sum e_p^2 - \left(\sum e_1^2 + \sum e_2^2\right)\right] / K}{\left(\sum e_1^2 + \sum e_2^2\right) / (n_1 + n_2 - 2K)}$

$$= \frac{[16{,}320 - (260.08 + 226.6)] / 2}{(260.08 + 226.6) / (35 + 30 - 2 \times 2)} \text{ with d.f.} = (K, n_1 + n_2 - 2K)$$

$$= \frac{(16{,}320 - 486.86)/2}{486.68/61} = \frac{7916.66}{7.9783} = 992.27$$

$F^* = 992.27$ with d.f. = (2, 61).

From the table value we see that [illegible]

Thus we see that [illegible] larger than the table value [illegible] level of significance. Hence the null hypothesis is to be rejected [illegible] level of significance. We may thus conclude that the MPC [illegible] cases differ significantly.

In particular if the pooled data are not given then

[illegible]

Here [illegible] with d.f. [illegible]

[illegible] 0.9730

From the table value we see that $F_{0.05}$ [illegible] = 3.84 (approx.)

We may thus conclude that the null hypothesis will be accepted as $F^* < F_{0.05}$ at 5% level of significance and hence there will be no difference in MPC in the two cases.

2.18 Extension of Linear Regression Model to Non-linear Relationships

In the simple linear regression model we consider a linear relation between two variables X and Y in the form $Y = \alpha + \beta X + u$. But in many situations this may not be the case. In economics we observe non-linear relationships among the variables.

Some of the most common forms of non-linear relations used in economics are given below.

(i) Demand curve with unit elasticity: $D = f(P)$, or $D = \frac{a}{P}$ where D represents quantity demanded and P denotes price.

(ii) Average cost curve: The traditional theory of U shaped cost curve may be approximated by a polynomial of third degree in output

$C = f(q)$ or $C_i = \alpha + \beta_1 q_i - \beta_2 q_i^2 + \beta_3 q_i^3 + u_i$

where C represents cost and q represents the level of output.

Now Average cost $\frac{C_i}{q_i} = \frac{\alpha}{q_i} + \beta_1 - \beta_2 q_i + \beta_3 q_i^2$ which is a U shaped curve.

(iii) Production function may be of the form $Q = f(K, L)$ or $Q_i = AL_i^{\alpha} K_i^{\beta} u_i$ where Q = level of output, L = labour employed, K = capital employed, α and β are two parameters. This type of production function is called Cobb-Douglas production function.

(iv) The production function may be of the form

$$Q_i = A\left[\delta K_i^{-\rho} + (1-\delta)L_i^{-\rho}\right]^{-\frac{1}{\rho}}$$

This type of production function is called CES production function. The symbols have their usual meaning.

Now to estimate the parameters of the non-linear [illegible] the non linear function [illegible] linear form and then we have [illegible] OLS method in the usual manner.

In order to find out the rate of change i.e. the slope of the [illegible] both linear and non-linear but converted into linear form with respect to the regressor as well as to find out the elasticity of the regressand with respect to the regressor we can consult the following table. The knowledge of these formulae also helps us to compare the various models.

Model	Equation	Slope $\left(\frac{dY}{dX}\right)$	Elasticity $\left(\frac{dY}{dX}\cdot\frac{X}{Y}\right)$
Linear	$Y = \alpha + \beta X$	β	$\beta\left(\frac{X}{Y}\right)^*$
Log linear	$\log Y = \alpha + \beta \log X$	$\beta\left(\frac{Y}{X}\right)$	β
Log linear	$\log Y = \alpha + \beta X$	$\beta\ (Y)$	$\beta\ (X)^*$
Linear log	$Y = \alpha + \beta \log X$	$\beta\left(\frac{1}{X}\right)$	$\beta\left(\frac{1}{Y}\right)^*$
Reciprocal	$Y = \alpha + \beta\left(\frac{1}{X}\right)$	$-\beta\left(\frac{1}{X^2}\right)$	$-\beta\left(\frac{1}{XY}\right)^*$
Log reciprocal	$\log Y = \alpha - \beta\left(\frac{1}{X}\right)$	$\beta\left(\frac{Y}{X^2}\right)$	$\beta\left(\frac{1}{X}\right)^*$

Note : * indicates that the elasticity is variable, depending on the value taken by X or Y or both. When no X and Y values are specified, in practice, very often these elasticities are measured at the mean values of these variables, namely, $\bar{X}$ and $\bar{Y}$.

Example 2.10. Estimate the investment function $I = f(r) = \alpha r^{\beta} u$ on the basis of the following information

$$n = 11,\ \sum_{i=1}^{n} Y_i = 12.2771,\ \sum_{i=1}^{n} X_i = 16.6729$$

$$\sum_{i=1}^{n} X_i^2 = 27.9605,\ \sum_{i=1}^{n} X_i Y_i = 15.1222,$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 3.4864,\ \sum_{i=1}^{n} (X_i - \bar{X})^2 = 2.6891,$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 4.8566,$$

where $Y = \log I$, $X = \log r$

Solution The regression function is given by $Y = \alpha X^{\beta} u$ which is non-linear form, where α and β are the two parameters whose values are to be estimated by the OLS method. Taking log on both sides we get

$\log Y = \log\alpha + \beta\log X + \log u$ or $Y^* = \alpha^* + \beta X^* + u^*$

where $\log Y = Y^*$, $\log\alpha = \alpha^*$, $\log X = X^*$ and $\log u = u^*$. This transformed function is linear in terms of logarithms.

The function $Y^* = \alpha^* + \beta X^* + u^*$ is of the form $Y = \alpha + \beta X + u$ and hence we apply the OLS method.

Now by the OLS method we can obtain the estimates of the parameters α^* and β. Thus we have

$$\text{(i)}\ \hat\beta = \frac{\sum_{i=1}^{n}(X_i^* - \bar X^*)(Y_i^* - \bar Y^*)}{\sum_{i=1}^{n}(X_i^* - \bar X^*)^2} = \frac{3.4864}{2.689} = 1.2965$$

$$\text{and } \hat\alpha^* = \bar Y^* - \hat\beta\bar X^* = 1.161 - (1.2965)\times(\text{[illegible]})$$
$$= [1.161 + 1.9851]$$
$$= \text{[illegible]}$$

$$\left[\text{Here } \bar Y^* = \sum_{i=1}^{n} Y_i^*/n = \frac{12.7721}{11} = 1.161 \text{ and } \bar X^* = \sum_{i=1}^{n} X_i^*/n = \frac{\text{[illegible]}}{11} = \text{[illegible]}\right]$$

$$\text{(ii)}\ R^2 = \frac{\hat\beta^2\sum_{i=1}^{n}(X_i - \bar X)^2}{\sum_{i=1}^{n}(Y_i - \bar Y)^2} = \frac{\hat\beta^2\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(Y_i - \bar Y)^2} = \frac{\hat\beta^2\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$$

$$= \frac{(1.2965)^2 \times 2.6891}{4.8566} = 0.93 \qquad R^2 \approx 0.93$$

$$\text{(iii)}\ \hat\sigma_u^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - \hat\beta^2\sum x_i^2}{n-2}$$

$$= \frac{4.8566 - (1.2965)^2 \times 2.6891}{11-2} = \frac{4.8566 - 4.5201}{9} = \frac{0.3365}{9} = 0.0373$$

$$\text{(iv)}\ Var(\hat\beta) = \frac{\hat\sigma_u^2}{\sum x_i^2} = \frac{0.0373}{2.6891} = 0.01387$$

$$SE(\hat\beta) = \sqrt{0.01387} = 0.1177$$

v. The regression results can now be written as follows:

$I^* = \alpha^* + b r^*$

or, $\log I = 3.0812 - 1.12965 \log r$ [illegible]
(0.1177)

or [illegible] where [illegible] antilog of 3.081 [illegible])

The results show that the constant [illegible] that the demand for investment is interest elastic. This [illegible] theory.

2.19 Problem of Prediction / Forecasting Relating to a Two-variable Linear Regression Model

Usually we do not differentiate between prediction and forecasting. We use the two interchangeably. But these two terms are not identical. Prediction, in fact, means an estimation of any event happening (in the past, present or future). On the other hand, forecasting is always associated with a time dimension, i.e., the future estimation for some specific future duration or over a period of time. All forecasts are predictions, but not all predictions are forecasts, as when we use regression to explain the relationship between two variables. Forecast implies time series and future while prediction does not. When we are interested to predict or forecast about future, we call the regression analysis as **historical regression**. With the help of regression analysis we can forecast about the future value on the basis of past and present information of the said variables (X and Y). In the context of forecasting we may also distinguish between *ex-ante forecast* and *ex-post forecast*. Ex-ante forecast is a forecast that uses information available at the time of forecast, whereas expost forecast is a forecast that uses information beyond the time at which the forecast is made.

Let us define a classical linear regression model given by $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots, n$ with the help of the pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. We estimate the relationship by the method of least squares. The estimated relationship is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. In case of time series data we write the regression equation as $Y_t = \alpha + \beta X_t + u_t$.

Now for some value of X (the independent variable), which is not in the sample, we may like to estimate the value of Y (the dependent variable).

The process of finding the value of the dependent variable from the estimated relationship for the known value of the independent variable not in the sample is called "Prediction".

Let us suppose that X_0 is the value of the independent variable not in the sample and we have to predict the value of Y when $X = X_0$. There are two types of prediction: (i) Point prediction, (ii) Interval prediction.

2.19.1 Point Prediction

When prediction is done in terms of a single value of the dependent variable, then it is called point prediction. We simply put $X = X_0$ in the estimated relationship and we get, $\hat{Y} = \hat{\alpha} + \hat{\beta}X_0 = \hat{Y}_0$.

Now [illegible] is the true value of the dependent variable Y when prediction is made and it is given by

$Y_0 = \alpha + \beta X_0 + u_0$, where u_0 is the corresponding value of the disturbance term.

So $E(Y_0) = \alpha + \beta X_0 + E(u_0)$

or $E(Y_0) = \alpha + \beta X_0$, $E(u_0) = 0$

Let us define the prediction error by $e_0 = Y_0 - \hat{Y}_0$ when we want to predict Y_0

and $e_0' = E(Y_0) - \hat{Y}_0$ is the prediction error when we want to predict $E(Y_0)$.

Mean of the Predictor

When we want to predict Y_0, then $e_0 = Y_0 - \hat{Y}_0$

Now $e_0 = Y_0 - \hat{Y}_0 = \alpha + \beta X_0 + u_0 - \hat{\alpha} - \hat{\beta} X_0$

or $e_0 = u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0$

or $E(e_0) = E(u_0) - E(\hat{\alpha} - \alpha) - E(\hat{\beta} - \beta) X_0 = 0$

where $E(u_0) = 0$, $E(\hat{\alpha}) = \alpha$, $E(\hat{\beta}) = \beta$

Hence $E(Y_0 - \hat{Y}_0) = 0$ i.e. $E(e_0) = 0$

or, $Y_0 - E(\hat{Y}_0) = 0$ or, $E(\hat{Y}_0) = Y_0$

So $\hat{Y}_0$ the OLS point predictor of Y_0 is unbiased.

Variance of the predictor Variance of the predictor is given by,

$\mathrm{Var}(\hat{Y}_0) = E[\hat{Y}_0 - E(\hat{Y}_0)]^2 = E[\hat{Y}_0 - Y_0]^2$ $[\because E(\hat{Y}_0) = Y_0]$

$= E[-e_0]^2 = E(e_0)^2$ $[\because e_0 = Y_0 - \hat{Y}_0]$

Now $E(e_0^2) = E[u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0]^2$

$$= E[u_0^2 + (\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 X_0^2 - 2(\hat{\alpha} - \alpha) u_0$$

$$- 2(\hat{\beta} - \beta) X_0 u_0 + 2(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) X_0]$$

$$= E(u_0^2) + E(\hat{\alpha} - \alpha)^2 + E(\hat{\beta} - \beta)^2 X_0^2 - 2E(\hat{\alpha} - \alpha) u_0$$

$$- 2X_0 E(\hat{\beta} - \beta) u_0 + 2X_0 E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$$

$$= \sigma_u^2 + \mathrm{var}(\hat{\alpha}) + X_0^2\, \mathrm{var}(\hat{\beta}) - 2\,\mathrm{cov}(\hat{\alpha}, u_0)$$

$$- 2X_0\, \mathrm{cov}(\hat{\beta}, u_0) + 2X_0\, \mathrm{cov}(\hat{\alpha}, \hat{\beta})$$

We know that, $\mathrm{var}(\hat{\alpha}) = \sigma_u^2 \left[\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right]$ and $\mathrm{cov}(\hat{\alpha}, u_0) = \mathrm{cov}(\hat{\beta}, u_0) = 0$ as the estimators of the parameters are independent of the disturbance term.

Now $\mathrm{cov}(\hat{\alpha}, \hat{\beta}) = E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right]$

[illegible]

$= -\bar{X}E(\hat{\beta} - \beta)^2 = -\bar{X}\,\mathrm{var}(\hat{\beta})$

So, $\mathrm{var}(\hat{Y}_0) = E(e_0^2)$

[illegible]

[illegible]

[illegible]

$$\text{Since } \mathrm{var}(\hat{\alpha}) = \sigma_u^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right] = \frac{\sigma_u^2}{n} + \bar{X}^2\frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} = \frac{\sigma_u^2}{n} + \bar{X}^2\,\mathrm{var}(\hat{\beta})$$

$$\mathrm{var}(\hat{Y}_0) = \sigma_u^2 + \frac{\sigma_u^2}{n} + \bar{X}^2\,\mathrm{var}(\hat{\beta}) + (X_0 \; \bar{X})\,\mathrm{var}(\hat{\beta}) \; \bar{X} \; \mathrm{var}(\hat{\beta})$$

or $\mathrm{var}(\hat{Y}_0) = E(e_0^2) = \sigma_u^2\left(1 + \frac{1}{n}\right) + (X_0 - \bar{X})^2\,\mathrm{var}(\hat{\beta})$

Except $\mathrm{var}(\hat{\beta})$ all the terms are constant and positive

So, $\mathrm{var}(\hat{Y}_0)$ is minimum when $\mathrm{var}(\hat{\beta})$ is minimum

We know that, $\mathrm{var}(\hat{\beta})$ is minimum when $\hat{\beta}$ is the OLS estimator of β. Hence $\mathrm{var}(\hat{Y}_0)$ is minimum when $\hat{Y}_0$ is the OLS point-predictor of Y_0. This is the BLUE property of OLS point predictor. It means that OLS point predictor of Y_0 i.e. $\hat{Y}_0$ is the best linear unbiased predictor of Y_0.

We now consider another case where we want to make a point prediction of $E(Y_0)$.

Here prediction error is defined as $E(Y_0) - \hat{Y}_0 = e_0$. We know that $Y_0 = \alpha + \beta X_0 + u_0$

$E(Y_0) = \alpha + \beta X_0$ because $E(u_0) = 0$

$e_0 = \alpha + \beta X_0 - \hat{\alpha} - \hat{\beta}X_0$

$= (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})X_0 = -(\hat{\alpha} - \alpha) - (\hat{\beta} - \beta)X_0$

$e_0^2 = (\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 X_0^2 + 2X_0(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$

or, $E(e_0^2) = E(\hat{\alpha} - \alpha)^2 + X_0^2 E(\hat{\beta} - \beta)^2 + 2X_0 E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$

$E(e_0^2) = E\left[\hat{Y}_0 - E(Y_0)\right]^2 = \mathrm{var}(\hat{\alpha}) + X_0^2\,\mathrm{var}(\hat{\beta}) + 2X_0\,\mathrm{cov}(\hat{\alpha}, \hat{\beta})$

Since var [illegible] $= \frac{\sigma_u^2}{\sum x_i^2}$ [illegible] and

and cov($\hat\alpha$, $\hat\beta$) [illegible] var($\hat\beta$)

[illegible] $\hat\beta$ [illegible] and [illegible]

[illegible] var($\hat\beta$)

Now putting the values of var($\hat\alpha$) and cov($\hat\alpha$, $\hat\beta$) in the above expression we get

E [illegible] var($\hat Y_0$) = var(e_0) [illegible]

Now var(e_0) = E [illegible]

$\frac{\sigma_u^2}{n}$ + [illegible] var($\hat\beta$) [illegible] var($\hat\beta$) [illegible] var($\hat\beta$) [illegible] var($\hat\beta$)

var(e_0) is minimum when var($\hat\beta$) is minimum. Now var($\hat\beta$) is minimum where $\hat\beta$ is the OLS estimate of β. So var(e_0) is minimised when $\hat Y_0$ is the OLS point predictor [illegible] $E(Y_0)$. This is the BLUE property of the OLS point predictor [illegible]

2.19.2 Test of Significance of Predictor and Interval Prediction

Case 1 We want to test the null hypothesis $H_0: Y_0 = A$ (some specified value) against the alternative hypothesis $H_1: Y_0 \neq A$ or $H_1: Y_0 > A$ or $H_1: Y_0 < A$. We use $\hat Y_0$ as the appropriate statistic of Y_0 because $\hat Y_0$ is the BLUE predictor of Y_0. Now $\hat Y_0 = \hat\alpha + \hat\beta X_0$. Since $\hat Y_0$ is a linear function of $\hat\alpha$ and $\hat\beta$ ($\hat\alpha$ and $\hat\beta$ are normally distributed) so $\hat Y_0$ is also normally distributed.

Since $E(e_0) = E[\hat Y_0 - Y_0] = 0$, $E(\hat Y_0) = Y_0$

and $\text{var}(\hat Y_0) = E[\hat Y_0 - Y_0]^2 = E(e_0^2)$

$= \sigma_u^2 + \text{var}(\hat\alpha) + \text{var}(\hat\beta)X_0^2 - 2X_0\,\text{cov}(\hat\alpha, \hat\beta)$

$$= \sigma_u^2 + \sigma_u^2\left[\frac{1}{n} + \frac{\bar X^2}{\sum_{i=1}^{n} x_i^2}\right] + X_0^2\frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} - 2X_0\bar X\frac{\sigma_u^2}{\sum x_i^2}$$ where $\text{cov}(\hat\alpha, \hat\beta) = -\bar X\,\text{var}(\hat\beta)$

and $\text{var}(\hat\beta) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$

[illegible]

$$\text{var}(\hat{Y}_0) = \sigma_u^2\left[1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^{n} x_i^2}(\bar{X} - X_0)^2\right]$$

This means that $\hat{Y}_0$ is normally distributed with mean [illegible] and variance

σ_u^2 [illegible] $\sum$ [illegible]. Now if σ_u^2 is unknown, it is to be replaced by its unbiased

estimator $\sum_{i=1}^{n} e_i^2/(n-2)$. The appropriate test statistic will be given by

$$t = \frac{\hat{Y}_0 - Y_0}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 + \frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}} \sim t_{n-2}$$

it follows a 't' distribution with $(n-2)$ degrees of freedom.

Nature of the test If the alternative hypothesis is $H_1 : Y_0 \neq A$, then the null hypothesis will be accepted at 5% level of significance if $-t_{0.025,\, n-2} \le t \le t_{0.025,\, n-2}$ and will be rejected otherwise.

If the alternative hypothesis is $H_1 : Y_0 > A$ then $H_0 : Y_0 = A$ will be accepted at 5% level of significance if t (observed) $\le t_{0.05,\, n-2}$ (table) and will be rejected otherwise.

If the alternative hypothesis is $H_1 : Y_0 < A$, then $H_0 : Y_0 = A$ will be accepted at 5% level of significance if

t (observed) $\ge -t_{0.05,\, n-2}$ (table), and will be rejected otherwise.

The rejection of the null hypothesis on the basis of the sample data implies the significance of Y_0.

It should be noted that the same procedure can be used for 1% level of significance

It can be seen that 100(1 − α)% confidence interval of Y_0 would be

$$\hat{Y}_0 \pm t_{\alpha/2,\,n-2}\sqrt{\frac{\sum e_i^2}{n-2}\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]}$$

where α usually takes the value 0.05 or 0.01.

Case 2. We want to test the null hypothesis $H_0: E(Y_0) = A$ against the alternative $H_1: E(Y_0) \neq A$ or $H_1: E(Y_0) > A$ or $H_1: E(Y_0) < A$.

Here we take $\hat{Y}_0$ as the statistic of $E(Y_0)$ because $\hat{Y}_0$ is the BLUE predictor of $E(Y_0)$.

Here also $\hat{Y}_0$ is normally distributed with mean $E(Y_0)$ and variance

$$\text{var}(\hat{Y}_0) = \frac{\sigma_u^2}{n} + (X_0 - \bar{X})^2\,\text{var}(\hat{\beta})$$

$$= \frac{\sigma_u^2}{n} + (X_0 - \bar{X})^2 \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} = \sigma_u^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]$$

So, $\hat{Y}_0$ is normally distributed with mean $E(Y_0)$ and variance $\sigma_u^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]$.

If σ_u^2 is unknown, it is to be replaced by its unbiased estimator $\sum_{i=1}^{n} e_i^2/(n-2)$. The appropriate test statistic under $H_0: E(Y_0) = A$ would be

$$\frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}}$$

which follows a t-distribution with $(n - 2)$ degrees of freedom.

Nature of the test

If $H_0: E(Y_0) = A$ is tested against the alternative $H_1: E(Y_0) \neq A$, then $H_0: E(Y_0) = A$ will be accepted at 5% level of significance if $-t_{0.025,\,n-2} \le t_{\text{observed}} \le t_{0.025,\,n-2}$ and will be rejected otherwise.

If the alternative hypothesis is H_1 : [illegible], then H_0 : [illegible] is to be rejected at the level of significance α if observed [illegible] table and is to be accepted otherwise.

If the alternative hypothesis is H_1 : [illegible], then H_0 : [illegible] will be accepted at the level of significance α if observed [illegible] table and will be rejected otherwise. The same test procedure is applicable at 1% level of significance.

It can be seen that 100(1 − α)% confidence interval of $E(Y_0)$ would be

$$\hat{Y}_0 \pm t_{\alpha/2} \sqrt{\frac{\sum e^2}{n-2}\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x^2}\right)}$$

where α usually takes the value 0.05 or [illegible]

Example 2.11. Following example 2.5.

(i) Find out the point predictor of Y_i when $X_i = 10$

(ii) It is claimed that when $X_i = 10$, $Y_i = 165$. Do you think that it is justified?

(iii) It is claimed that when $X_i = 10$, $E(Y_i) = 155$. Do you think that the claim is justified?

Solution

(i) We have to find out the point predictor of Y_i when $X_i = 10$. We know that the point predictor of Y_i is,

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ (where $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$. See Ex. 2.5)

$= -3.505 + 0.494 \times 10$

$= -3.505 + 4.94 = 1.435$

$\hat{Y}_i = 1.435$

So, point predictor of Y_i is $\hat{Y}_i = 1.435$ when $X_i = 10$

(ii) We have to examine whether $Y_i = 165$ when $X_i = 10$ is justified or not. We have to test the null hypothesis $H_0 : Y_0 = 165$, against the alternative $H_1 : Y_0 \neq 165$

The appropriate test statistic is given by,
$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\dfrac{1}{n}+\dfrac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}} \sim t_{n-2}$$

Since $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$ and $X_0 = 10$,

then $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0 = -3.505 + 0.494 \times 10 = 1.435$

$\hat{Y}_0 - Y_0 = 1.435 - 165.000 = -163.565$

and $\sqrt{\ldots}$ [illegible]

$= \sqrt{\ldots}$ [illegible]

$= \sqrt{\ldots}$ [illegible]

$= \sqrt{\ldots}$ [illegible]

$= 4.69$

Since $\frac{\sum X_i}{n} = \ldots$ [illegible]

and $\sum x_i^2 = 2\ 54$ [illegible]

Now t (observed) $= \dfrac{\hat{Y}_0 - Y_0}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 + \dfrac{1}{n} + \dfrac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}}$

$= \dfrac{-\ldots}{4.69} = -115.438$

$t = -115.438$

Here we see that t (observed) $= -115.438$ which does not lie in the interval [illegible] i.e., in the interval −2.101 and 2.101 and hence the null hypothesis $H_0: Y_0 = 165$ is rejected for the given sample at 5% level of significance. So, Y_0 [illegible] is not justified when $X_0 = 10$.

(iii) We have to examine whether $E(Y_0) = 155$ when $X_0 = 10$.

We have to test the null hypothesis $H_0: E(Y_0) = 155$ against the alternative $H_1: E(Y_0) \neq 155$.

The statement $E(Y_0) \neq 155$ when $X_0 = 10$ will be justified if the null hypothesis $H_0: E(Y_0) = 155$ is rejected.

The appropriate test statistic would be

$$t = \frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}} \sim t_{n-2}$$

Here [illegible] $E(Y_0) = 155$

[illegible]

Now [illegible]

$X_0 = 10$, $\bar{X} = 9.3$, $\sum x_i^2 = 215.4$

$$\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}$$

$$\sqrt{1.908\left[\frac{1}{20}+\frac{(10-9.3)^2}{215.4}\right]}$$

$$= \sqrt{1.908[0.05 + 0.00221]} = \sqrt{1.908 \times 0.05221} = \sqrt{0.0996} = 0.3156$$

Now t (observed) $= \dfrac{\hat{Y}_0 - E(Y_0)}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\dfrac{1}{n}+\dfrac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}} = \dfrac{-153.565}{0.3156} = -486.58$

t (observed) $= -486.581$

Now on the basis of the given sample the null hypothesis $H_0: E(Y_0) = 155$ will be accepted at 5% level of significance if $-t_{0.025,n-2} \le t \le t_{0.025,n-2}$ and will be rejected otherwise.

Here $t_{0.025,n-2} = t_{0.025,18} = 2.101$

So, the observed $t = -486.581$ does not lie in the interval -2.101 and 2.102 and hence the null hypothesis will be rejected. So, $E(Y_0) = 155$ when $X_0 = 10$ is not justified.

Example 2.12. Consider the following regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i is normally distributed with mean zero and variance σ_u^2 (unknown). We have the following data

X	2	3	1	5	9
Y	4	7	3	9	7

(i) Estimate α and β.

(ii) Test whether α and β are significant or not at 5% level of significance.

(iii) Calculate R^2

[illegible]

Solution: [illegible] estimate the regression parameters α and β [illegible] and [illegible] α and β. We know that

[illegible]

Calculations for α, β and R^2

X	Y	XY	X^2	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
[illegible]	4	8	4	2	[illegible]	4	4	4
[illegible]	[illegible]	21	9	[illegible]	[illegible]	1	1	[illegible]
5	7	3	1	3	1	9	[illegible]	[illegible]
1	9	45	25	1	1	1	[illegible]	[illegible]
6	[illegible]	63	36	5	1	25	1	5
$\sum_{i=1}^{n} X_i = 20$	$\sum_{i=1}^{n} Y_i = 30$	$\sum_{i=1}^{n} X_iY_i = 40$	$\sum_{i=1}^{n} X_i^2 = 20$	$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 0$	$\sum_{i=1}^{n} x_i^2 = 40$	$\sum_{i=1}^{n} y_i^2 = 24$	$\sum_{i=1}^{n} x_iy_i = 20$

Here $\bar{X} = \sum_{i=1}^{n} X_i / n = \frac{20}{5} = 4$ and $\bar{Y} = \sum_{i=1}^{n} Y_i / n = \frac{30}{5} = 6$

Now $\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{20}{40} = 0.5$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 6 - 0.5 \times 4 = 6 - 2 = 4$

$\therefore \hat{\alpha} = 4, \hat{\beta} = 0.5$

Thus the OLS estimators of α and β are $\hat{\alpha} = 4$ and $\hat{\beta} = 0.5$

The estimated regression line is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $\hat{Y} = 4 + 0.5X$

Now we have to find out $Var(\hat{\alpha})$ and $Var(\hat{\beta})$. We know that

$$Var(\hat{\alpha}) = \sigma_u^2 \frac{\sum_{i=1}^{n} X_i^2}{\left(n \sum_{i=1}^{n} x_i^2\right)} \quad \text{and} \quad Var(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2$$

But here σ_u^2 is not known and hence it is to be replaced by its unbiased estimator

$$\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2)$$

[illegible]

[illegible]

[illegible]

Now Var(α̂) = σ_u^2 [illegible] = 4.67 × [illegible] 5×40 [illegible]

$$Var(\hat{\alpha}) = 2.802 \text{ and } SE(\hat{\alpha}) = \sqrt{Var(\hat{\alpha})} = \sqrt{2.802} = 1.674$$

$$\text{Similarly } Var(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2 = \frac{4.67}{40} = 0.11675$$

$$Var(\hat{\beta}) = 0.11675 \text{ and } SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{0.11675} = 0.3416$$

(ii) **Test for $\hat{\alpha}$ and $\hat{\beta}$:**

a. **Test for β** We have to test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$. The appropriate test statistic would be

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-2}$$

The null hypothesis will be accepted at 5% level of significance if $-t_{0.025, n-2} < t < t_{0.025, n-2}$ and will be rejected otherwise.

Here we see that,

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sqrt{\left(\sum_{i=1}^{n} e_i^2 / (n-2)\right) \Big/ \sum_{i=1}^{n} x_i^2}} = \frac{0.5}{0.3416} = 1.4637$$

t (observed) = 1.4637

But from table value $t_{0.025, n-2} = t_{0.025, 5-2} = t_{0.025, 3} = 3.182$

Here we see that the t (observed) = 1.4637 lies in the interval −3.182 and 3.182 and hence the null hypothesis is accepted at 5% level of significance.

So, $\hat{\beta}$ is insignificant at 5% level, significant only when the null hypothesis is rejected.

(b) Test for $\hat{\beta}$. If we test the null hypothesis H_0 [illegible] 0 against [illegible] then the null hypothesis will be accepted at 5% [illegible] and will be rejected otherwise [illegible]

[illegible]

From [illegible] values [illegible]

and [illegible] $= \frac{4}{1.674} =$ [illegible]

[illegible]

Thus we see that t (observed) [illegible] 1895 lies in the interval [illegible] and [illegible] 82 and [illegible], and hence the null hypothesis is accepted at 5% level of significance. This means that it is also insignificant at 5% level of significance.

(c) Now we have to calculate the value of R^2

$$\text{Since we know that } R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{(0.5)^2\, 40}{24} = \frac{10}{24} = 0.4167 \approx 0.42$$

$$R^2 = 0.42 = \frac{42}{100} = \frac{\text{Explained variation}}{\text{Total variation}}$$

This suggests that 42 percent of the variations in the sample observations of Y can be attributed to the variations of the fitted value of Y ($\hat{Y}_i$) or we can say that our regression line fits the given data not very well.

From the above results we can write our regression results as follows

$$\hat{Y} = 4 + 0.5X \quad R^2 = 0.42$$

$$SE \quad (1.674) \; (0.3416)$$

Alternatively, $\hat{Y} = 4 + 0.5X$, $R^2 = 0.42$

t ratios (2.3895) (1.4637)

(v) We have to find out the point predictor of Y when $X = 10$

The point predictor of Y is given by $\hat{Y} = \hat{\alpha} + \hat{\beta}X$

$= 4 + 0.5 \times 10 = 4 + 5 = 9$

Point predictor $\hat{Y} = 9$ when $X = 10$

Example 2.13. Following the data given in Example 2

(i) estimate the regression parameters (assuming a linear regression equation of the form $Y_i = \alpha + \beta X_i + u_i$ where $u_i \sim N(0, \sigma_u^2)$ (unknown)

(ii) Calculate R^2.

(iii) Comment about the [illegible] if [illegible] expenditures are increased [illegible].

(iv) Find out OLS [illegible] when $X = $ [illegible]

(v) Find out OLS [illegible] of the [illegible] predicted value of Y when X [illegible]

Solution

Table for calculation

Month	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2	$e = Y - \hat{Y}$	e^2
1	1	3	−2	−1.6	3.2	4	2.56	0.8	0.64
2	2	4	−1	−0.6	0.6	1	0.36	0.6	0.36
3	3	2	0	−2.6	0	0	6.76	−2.6	6.76
4	4	6	1	1.4	1.4	1	1.96	0.2	0.04
5	5	8	2	3.4	6.8	4	11.56	1.0	1.00
Total	$\sum X = 15$	$\sum Y = 23$	$\sum x = 0$	$\sum y = 0$	$\sum xy = 12$	$\sum x^2 = 10$	$\sum y^2 = 23.20$	$\sum e = 0$	$\sum e^2 = 8.8$

Here $n = 5$ as we have data for five months.

Now $\bar{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$, $\bar{Y} = \frac{\sum Y}{n} = \frac{23}{5} = 4.6$

(i) We have to estimate the regression parameters α and β. Let $\hat{\alpha}$ and $\hat{\beta}$ be the OLS estimators (predictors here) of α and β.

We know that $\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$\hat{\beta} = \frac{12}{10} = 1.2$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 4.6 - 1.2 \times 3 = 4.6 - 3.6 = 1$

Therefore the estimated (predicted) sample regression equation is $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i = 1.0 + 1.2X_i$

In the table $e_i = Y_i - \hat{Y}_i$ can be obtained for different values of X. Since $e_i = Y_i - 1.0 - 1.2X_i$.

When $X_i = 1$ and $Y_i = 3$, $e_i = 3 - 1.0 - 1.2 \times 1 = 3.0 - 2.2 = 0.8$

$X_i = 2$ and $Y_i = 4$, $e_i = 4 - 1.0 - 1.2 \times 2 = 4 - 3.4 = 0.6$

and in this way other e_i values are calculated.

(ii) We know that $R^2 = \dfrac{ESS}{TSS} = \dfrac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \dfrac{(1.2)^2 \times 10}{23.20} = \dfrac{14.4}{23.20} = 0.620$

Since $\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$ i.e., TSS = ESS + RSS ∴ $R^2 = 0.620$

[illegible]

[illegible]

[illegible]

[illegible] ₹ [illegible]

We have to find out 95% confidence interval of the (predicted) value of [illegible] when X = ₹ 600

We know that at 95% [illegible] (α = 0.05) [illegible] confidence interval of the [illegible] predicted [illegible]

[illegible]

Here we have $\hat{Y}_0$ = [illegible] when $X = X_0 = 6$

and $\sum_{i=1}^{n} e_i^2 = 8.8$, $n = 5$, $\sum_{i=1}^{n} x_i^2 = 10$, $\bar{X} = 3$, $\frac{\sum_{i=1}^{n} e_i^2}{(n-2)} = \frac{8.8}{3} = 2.93$

and $t_{\alpha/2, n-2} = t_{0.025, 3} = 3.182$ [From table value]

So, 95% confidence interval of the predicted value of sales revenue corresponding to advertising expenditure of ₹ 600 would be

$$\approx 20 \pm 3.182\sqrt{2.93\left[1 + \frac{1}{5} + \frac{(6-3)^2}{10}\right]}$$

or, ≈ 20 ± 7.091 or 0.309 and 16.09

⇒ ₹ 309 and ₹ 1609]

Thus 95% confidence interval of predicted sales revenue (Y_0) corresponding to advertising expenditure of ₹ 600 would be ₹ 309 and ₹ 1609]

(vi) 95% [illegible] when α = 0.05] confidence interval of expected sales revenue $(E(\hat{Y}_0))$ when advertising expenditures are ₹ 600 would be

$$\hat{Y}_0 \pm t_{\alpha/2, n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]}$$

[Here α = 0.05, $t_{\alpha/2, n-2} = t_{0.025, 3} = 3.182$ [from table value]

$n = 5$, $X_0 = 6$, $\sum x_i^2 = 10$, $\bar{X} = 3$, $\sum_{i=1}^{n} e_i^2 / (n-2) = 2.93$

The prediction is still given by

$\hat{Y}_0 = 1.0 + 1.2X$, so that when $X =$ [illegible] $= 6$, $\hat{Y}_0 = 1.0 + 1.2 \times 6 = 8.20$

[illegible]

or, $8.20 \pm 3.182 \times 1.798$

or, 8.20 ± 5.72 or, 2.48 and 13.92

95% confidence interval for the average sales $E(Y_0)$ corresponding to advertising expenditures ₹ 600 would be ₹ 2480 and ₹ 13920.

It should be noted that this confidence interval is narrower than the one we obtained for $\hat{Y}_0$.

Example 2.14 The following table (Table 2.6) gives data on the level of education (number of years of schooling), the mean hourly wages earned by the people at each level of education and the number of people at the stated level of education.

Table 2.6. Mean Hourly wage by Education

Years of schooling (X)	Mean hourly wage in \$ (Y)	Number of people
6	4.4567	3
7	5.7700	5
8	5.9787	15
9	7.3317	12
10	7.3182	17
11	6.5844	27
12	7.8182	218
13	7.8351	37
14	11.0223	56
15	10.6738	13
16	10.8361	70
17	13.6150	24
18	13.5310	31
Total		528

(i) Assuming a linear regression line of the form $Y_i = \alpha + \beta X_i + u_i$ [$u_i \sim N(0, \sigma_u^2)$], find the OLS estimators of α and β

(ii) Find var $(\hat{\alpha})$ and var $(\hat{\beta})$

(iii) Find $SE(\hat{\alpha})$ and $SE(\hat{\beta})$

(iv) Find R^2

(v) Find $\sum_{i=1}^{n} e_i^2$

(vi) Predict/Forecast about the mean hourly wage when the level of education (years of schooling) is 20

[illegible]

[illegible]

[illegible]

[illegible]

Calculations for the Regression

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	6	6	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	9	[illegible]	[illegible]	9	[illegible]	[illegible]
5	[illegible]	10	[illegible]	[illegible]	4	[illegible]	[illegible]
6	[illegible]	11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	12	0	[illegible]	0	[illegible]	[illegible]
8	[illegible]	13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	14	2	[illegible]	4	[illegible]	[illegible]
10	[illegible]	15	3	[illegible]	9	[illegible]	[illegible]
[illegible]	[illegible]	16	4	[illegible]	16	[illegible]	[illegible]
12	[illegible]	17	5	4.9403	25	[illegible]	[illegible]
13	[illegible]	18	6	4.8563	36	[illegible]	[illegible]
Total	$\sum_{i=1}^{n} Y_i$ = [illegible]	$\sum X_i$ = 156	$\sum_{i=1}^{n} x_i$ = 0	$\sum_{i=1}^{n}$ = 0	$\sum x_i^2$ = 182	$\sum$ = [illegible]	$\sum$ [illegible]

Note $n = 13$, $\bar{Y} = \sum Y_i / n =$ [illegible] $= 8.6747$

and $\bar{X} = \sum_{i=1}^{n} X_i / n = \frac{156}{13} = 12$, $\sum_{i=1}^{n} x_i^2 = (6)^2 + (7)^2 +$ [illegible]

$$\beta = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \frac{131.7856}{182.0} = 0.7240967$$

and $\alpha = \bar{Y} - \beta\bar{X} = 8.6747 - 0.7240967 \times 12 = -0.01445$

The estimated regression equation is

$\hat{Y} = \alpha + \beta X$ or $\hat{Y} = -0.01445 + 0.7240967\,X_i$

$e_i = Y_i - \hat{Y}_i = Y_i + 0.01445 - 0.7240967\,X_i$

Now different values of e_i can be obtained by taking different pairs of values of X_i and Y_i

When [illegible] 8.456 and [illegible] then [illegible] the above [illegible] we get [illegible] similarly other values of x are obtained.

Now $\sum_{i=1}^{n}$ [illegible]

and $\sum_{i=1}^{n}$ [illegible]

(i) OLS estimates of α and β would be

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{131.7856}{182} = 0.7240967$$

$$\text{and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 8.6742 - 0.7240967 \times [illegible] = 0.01445$$

$$\hat{\alpha} = 0.0144 \text{ and } \hat{\beta} = 0.7240$$

(ii) We have to calculate the values of var ($\hat{\alpha}$) and var ($\hat{\beta}$)

We know that $\text{var}(\hat{\alpha}) = \dfrac{\sigma_u^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}$ Here σ_u^2 is unknown and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$$\text{var}(\hat{\alpha}) = \frac{\hat{\sigma}_u^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} = \frac{0.8936 \times 2054}{13 \times 182} = \frac{1835.4544}{2366} = 0.7757$$

$$\text{var}(\hat{\alpha}) = 0.7757$$

Again, $\text{var}(\hat{\beta}) = \dfrac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$ Here σ_u^2 is unknown and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$$\text{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \frac{0.8936}{182} = 0.004910 \qquad \text{var}(\hat{\beta}) = 0.004910$$

[illegible]

and [illegible]

[illegible] and [illegible]

Since [illegible] $\frac{\text{ESS}}{\text{TSS}}$ [illegible]

Since $\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$ or TSS = ESS + RSS

$$\sum_{i=1}^{n} y_i^2 = (0.7240)^2 \, [illegible] + 9.83017 \qquad \sum e_i^2 = 9.83017$$

$$= 94.4000 + 9.83017 \quad [illegible]$$

(iii) Note $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (0.79406)^2 + (0.853117)^2 + \dots + [illegible]$

$$= 9.83017 \qquad \sum_{i=1}^{n} e_i^2 = 9.83017$$

(iv) We have to forecast/predict about the mean hourly wage rate when the level education (years of schooling) is 20

Since the estimated regression equation is $\hat{Y}_i = -0.0144 + 0.7240 X_i$

The point predictor of Y is given by $\hat{Y} = \hat{\alpha} + \hat{\beta} X$

When $X = X_0$, $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$

So, when the level of education is $X_0 = 20$, the mean hourly wage rate would be

$\hat{Y}_0 = -0.0144 + 0.7240 \times 20 = 14.4656$

So, mean hourly wage rate would be $ 14.4656 when years of schooling increase to 20

(v) We have to construct 95% confidence interval for the point predictor of hourly wage rate ($\hat{Y}_0$) when the level of education becomes $X_0 = 20$. This confidence interval would be

$$\hat{Y}_0 \pm t_{0.025,\, n-2} \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 + \frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}$$

When $X = X_0 = 20$, $Y_0 = \hat{Y}_0 = -0.0144 + 0.7240 \times 20 = 14.4656$

[illegible] 1 (Table value) [illegible]

Here $n = 13$ [illegible] $\sum_{i=1}^{n}$ [illegible] $\sum_{i=1}^{n}$ [illegible]

Thus [illegible] 95% [illegible] confidence interval of [illegible] would be

$$Y_0 \pm t_{0.025,n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum x_i^2}\right]}$$

or $14.4656 \pm 2.201\sqrt{0.8936\left[1+\frac{1}{13}+\frac{(12-20)^2}{182}\right]}$

or $14.4656 \pm 2.201\sqrt{0.8936\left[1+\frac{1}{13}+\frac{64}{182}\right]}$

or $14.4656 \pm 2.201\sqrt{0.8936[1+0.07692+0.35164]}$

or $14.4656 \pm 2.201\sqrt{0.8936 \times 1.42856}$ or $14.4656 \pm 2.201\sqrt{1.27656}$

or $14.4656 \pm 2.201 \times 1.12984$ or, 14.4656 ± 2.4868

i.e., 11.9788 and 16.9524

95% confidence interval of hourly wage rate would be \$ 11.9788 and \$ 16.9524 when the level of education (years of schooling) is 20

(vi) We have to construct 95% confidence interval of expected mean hourly wage rate when the level of education is $X_0 = 20$ (years of schooling)

When $X = X_0 = 20$ $E(Y/X_0 = 20)$ can be obtained as $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0 = -0.0144 + 0.7240 \times 20 = 14.4656$

Thus 95% confidence interval of $E(Y/X_0)$ when $X_0 = 20$ would be

$$\hat{Y}_0 \pm t_{0.025,n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}$$

Here $\hat{Y}_0 = 14.4656$, $n = 13$, $t_{0.025,n-2} = t_{0.025,11} = 2.201$ (Table value)

$\bar{X} = 12$, $X_0 = 20$ and $\sum_{i=1}^{n} x_i^2 = 182$, $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

Now, Y_0 [illegible]

$$= 14.4656 \pm 2.201\sqrt{0.8946\left[\ldots\right]}$$

or $14.4656 \pm 2.201\sqrt{0.8946 \; [illegible]}$ or $14.4656 \pm 2.201\sqrt{0.4829}$

or $14.4656 \pm 2.201 \times [illegible]$ or 14.4656 ± 1.3620

i.e., 13.1036 and 15.8276

Thus, 95% confidence interval of expected mean hourly wage rate corresponding to the level of education (year of schooling) X_0 [illegible] would be $ [illegible] and $ [illegible]

Example 2.15. The following table (Table 2.7) shows consumption expenditure and income (in billions of $) of a country over the period 2007–2018.

Table 2.7 Consumption expenditure and income of a country (in billions of $)

Year	Consumption expenditure (Y)	Income (X)
2007	[illegible]	[illegible]
2008	[illegible]	[illegible]
2009	[illegible]	[illegible]
2010	[illegible]	[illegible]
2011	[illegible]	[illegible]
2012	[illegible]	[illegible]
2013	[illegible]	[illegible]
2014	[illegible]	[illegible]
2015	434	[illegible]
2016	[illegible]	[illegible]
2017	[illegible]	[illegible]
2018	[illegible]	[illegible]

From the data given in the table we have the following results

$\hat{Y}_i = [illegible] + 0.7[illegible]\, X_i$ $\quad R^2 = 0.99[illegible]$, $\quad \hat{\sigma}_u^2 = \sum e_i^2/(n-2) = 28[illegible]$

$(5.39) \quad (10.0[illegible])$

$\bar{X} = 498, \quad n = 12$

$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} x_i^2 = 51,45[illegible]$ [illegible]

X_0 = Income in the year 2025 = $850 billion

(i) Forecast about consumption expenditure of the country for the year 2025 if income in that year increases to $850 billion.

(ii) Construct 95% confidence interval of the consumption expenditure. Predicted (forecasted) for the year 2025.

Solution

$\hat{Y}$ [illegible] $= 31.76 + 0.71\,X_t$ or, when $X_t = 850 + \hat{Y}_n$

$31.76 + 0.71 \times 850 = 635.26 \approx \$635$ billion

[illegible] of the [illegible] be found [illegible] the [illegible] for the year [illegible] be

$$[illegible]$$

or $\quad 635 \pm t$ [illegible] $\sqrt{285.61\left[1+\frac{1}{12}+\frac{123.904}{151.482}\right]}$

or $\quad 635 \pm 2.228\sqrt{285.6}$ [illegible]

or $\quad 635 \pm 2.228\sqrt{534.5885}$ [illegible]

or $\quad 635 \pm 2.228 \times 23.1$ [illegible]

or $\quad 635 \pm 51.4884$

or $\quad 583.5116$ and 686.4884

Thus 95% confidence interval of predicted consumption expenditure of the country for the year 2025 would be $583.5116 billion and $686.4884 billion.

EXERCISE

1. In a simple linear regression model, $Y_t = \alpha + \beta X_t + u_t$ for $t = 1, 2, \ldots, n$ why do we insert the random disturbance term u_t?
2. State and explain the assumptions of a classical linear regression model (CLRM).
3. In a simple linear regression model of the form $Y_t = \alpha + \beta X_t + u_t$, $t = 1, 2, \ldots, n$ how can you estimate the regression parameters α and β?
4. Describe briefly the method of moments used in estimating the regression parameters in a two variable linear regression model.
5. Describe briefly the method of least squares used in estimating the regression parameters relating to a two variable linear regression model.
6. How can you estimate a linear function (two variable) whose intercept is zero?
7. How can you estimate the elasticities from an estimated regression line?
8. State and prove the properties of the least squares estimators relating to a two variable linear regression model (CLRM).
9. Show that in a classical linear regression model the estimated regression parameters are unbiased.
10. Determine the mean and variance of $\hat{\alpha}$ and $\hat{\beta}$ relating to a model $Y_t = \alpha + \beta X_t + u_t$ for $t = 1, 2, \ldots, n$.

estimate α and β and calculate estimates of variances of your estimates. Estimate the conditional mean value of Y corresponding to a value of X fixed at X [illegible]

27. The following table shows the investment expenditure and the [illegible] interest rate [illegible] the ten year period

Year	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Investment	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Interest	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Test the hypothesis that investment is interest elastic by fitting a regression line to the above data and indicating the relevant levels of significance.

28. Given the following data

$\sum x_i y_i$ [illegible] $\sum x_i^2 =$ [illegible] $\sum y_i^2 =$ [illegible] [illegible] estimate the parameters

in the model $Y_i = \alpha + \beta X_i + u_i$ and test the hypothesis $H_0: \beta$ [illegible] against the alternative (i) $\beta \neq 1.5$, (ii) $\beta > 1.5$, (iii) $\beta < 1.5$

29. The true relationship between X and Y in the population is given by $Y = 2 + 3X + u$. Suppose the value of X in the sample of 10 observations are [illegible] 3, [illegible] 15. The values of the disturbances are drawn at random from a normal population with zero mean and constant variance

$u_1 = 0.464$, $u_2 =$ [illegible], $u_3 = 1.48$, $u_4 =$ [illegible], $u_5 =$ [illegible], $u_6 =$ [illegible], $u_7 =$ [illegible], $u_8 = 0.64$, $u_9 = 0.18$ and $u_{10} = -1.37$

(i) Present the 10 observed values of X and Y

(ii) Estimate the least squares estimates of the regression coefficients and their standard errors.

(iii) Obtain the predicted value of Y for X = 12

30. The following data gives the production of coal and the number of wage earners in the coal industry

Output (million tonnes)	[illegible]	[illegible]	211.5	208.9	207.4	205.3	196.8	192.1	183.2	[illegible]
Number of workers ('000 s)	706.2	703	701.8	699.1	697.4	795.3	692.7	630.2	[illegible]	[illegible]

(i) Estimate the production function (linear) of coal

(ii) Find average and marginal productivity of labour.

(iii) Estimate t-ratios and test their significance

31. The following are data on

Y = Quit rate per 100 employees in manufacturing

X = unemployment rate

The data are for the United States and cover the period 1960-1972

Year	Y	X	Year	Y	X
1960	1.3	6.2	1966	2.6	3.2
1961	1.2	7.8	1967	2.3	3.6
1962	1.4	5.8	1968	2.5	3.3
1963	1.4	5.7	1969	2.7	3.3
1964	1.5	5.0	1970	2.1	5.6
1965	1.9	4.0	1971	1.8	6.8
			1972	2.2	5.6

11. [illegible] with reference to the CLRM.

12. [illegible]

13. [illegible] the random disturbance term in the [illegible]

14. Show that $\sum$ [illegible] is an unbiased estimator of the variance of the disturbance term u_i in the model $Y_i = \alpha + \beta X_i + u_i$.

15. What is a maximum likelihood estimator (MLE)? Show that in the model [illegible]

[illegible] is the MLE of [illegible]

[illegible] is the MLE of [illegible]

(iii) $\sum$ [illegible] is the MLE of σ_u^2

16. Show that the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are such that

(i) [illegible] (ii) [illegible]

17. Describe the testing procedure of the significance of the regression coefficients of the model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots, n$.

18. What is meant by goodness of fit of the correlation coefficient R^2?

19. Show that Total sum of squares = Explained sum of squares + Unexplained sum of squares.

20. What is coefficient of determination? Show that it lies between 0 and 1 and hence show that the value of correlation coefficient between two variables lies between [illegible] and [illegible]

21. How can you formally write the regression results of the regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i (for $i = 1, 2, \ldots, n$) satisfies all the properties of CLRM?

22. How can you use the analysis of variance in the simple classical linear regression model?

23. What is the meaning of the term Prediction? How can you incorporate the term in the CLRM? Distinguish between point prediction and interval prediction in this regard.

24. Show that the OLS point predictor in the CLRM satisfies the [illegible] property.

25. The following sums where obtained from 16 pairs of observations on X and Y: [illegible] 526, $\Sigma X_i^2 = 657$, $\Sigma X_i Y_i = 492$, $\Sigma Y_i = 63$, $\Sigma X_i = 96$. Estimate the parameters of the model $Y = \alpha + \beta X + u$, and R^2.

Test the hypothesis that $\beta = 2.0$.

26. A sample of 20 observations corresponding to the regression model $Y = \alpha + \beta X + u$ gave the following data:

$\Sigma Y_i = 2$[illegible]9, $\Sigma (Y_i - \bar{Y})^2 = 86.9$, $\Sigma (X_i - \bar{X})(Y_i - \bar{Y}) = 106.4$, $\Sigma X = $ [illegible]86.2, $\Sigma (X_i -$ [illegible]

[illegible]

[illegible]

[illegible]

32. [illegible] function [illegible] What is the marginal [illegible] the marginal propensity [illegible] functions

Year	[illegible]	Consumption [illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

33. A random sample of ten families had the following income and food expenditure per week.

Families	A	B	C	D	E	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Family Income	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Family Expenditure	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Estimate the regression line of food expenditure on income and interpret your results.

34. The following results have been obtained from a sample of [illegible] observations on the value of sales [illegible] of a firm and the corresponding prices [illegible]

[illegible]

(i) Estimate the regression line of sales on price and interpret the results.

(ii) What is the part of the variation in sales which is not explained by the regression line?

(iii) Estimate the price elasticity of sales.

35. The following table gives the quantities of commodity Z bought in each year [illegible] and the corresponding prices.

Year	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Quantity (in tons)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Price in ₹	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

(i) Estimate the linear demand function for commodity Z.

(ii) Calculate the price elasticity of demand.

(iii) Forecast the demand at the mean price of the sample.

(iv) Forecast the demand at P = [illegible]

36. A sample of 20 observations [illegible] is to be used for estimating the linear function [illegible] The first 10 observations yield the following results:

$$\bar{X} = [illegible], \quad \bar{Y} = [illegible], \quad \sum_{i=1}^{10} [illegible] = [illegible], \quad \sum_{i=1}^{10} [illegible] = [illegible]$$

$$\sum_{i=1}^{10} [illegible] = [illegible]$$

36. 10 subsequent pairs of values of X and Y yield [illegible]

[illegible]

(iii) [illegible]

37. The following table [illegible] data on the quantity [illegible] commodity [illegible] and its price.

Year	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Quantity [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	60	20
[illegible] price [illegible] $ per unit)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

(i) Test the hypothesis that the quantity supplied and price are [illegible] by estimating the export supply function [illegible]

(ii) Show that β is a part of the price [illegible] of supply [illegible] a numerical value for the latter.

(iii) [illegible] price [illegible] becomes [illegible] and [illegible] then [illegible] the export volume of the commodity in these years.

38. The following table includes the total cost and the [illegible] output of firm A over a ten year period.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quantity (X) (000 units)	40	42	46	55	66	70	88	100	120	140
Total cost (Y) (000 dollars)	150	140	160	170	150	162	185	165	190	185

(i) Estimate the linear cost function $Y = \alpha + \beta X$

(ii) Find the AVC, MC and AC and plot them roughly on a graph.

39. The total investment function for the economy as a whole is assumed to be of the form

$I = \alpha r^{\beta} e^{u}$

where I = investment, r = rate of interest

The following sample is given

I ($ billion)	9.0	5.5	8.5	4.0	3.5	2.5	3.0	1.5	1.2	1.8	1.5
r (percent)	2	3	2	4	5	6	4	6	8	7	9

(i) Estimate the parameters of the investment function by OLS.

(ii) Test the statistical significance of the coefficients at 5% level of significance.

(iii) Construct a 95% confidence interval for β.

(iv) Find the value of R^2 and interpret the result.

40. a. Write down the assumptions essential for each of the following tasks.

i. Proving that the OLS estimators are unbiased

[illegible]

[illegible]

[illegible]

[illegible]

41. Consider the following estimated regression equation: $\hat{Y}$ [illegible] with estimated standard error of $\hat{\beta}$ [illegible] and 0.[illegible]. It has further been given that [illegible]

[illegible] and $\sum$ [illegible] 4895

Find out the following:

(a) Sample size (n) [illegible] The estimated intercept [illegible] Total sum of squares

(d) Residual sum of squares (RSS) (e) Estimated error variance ($\hat{\sigma}_u$)

42. (a) In a two variable regression model, show that $R^2 = \frac{ESS}{TSS}$

(b) Explain why random error term is introduced in an econometric model.

(c) Which one would you consider to be CLRM? Give reasons.

(i) $Y_i = \beta_1 + \beta_2 X_i$ (ii) $Y_i = \beta_1 + \sqrt{\beta_2} X_i$ (iii) $Y_i = \beta_1 + \beta_2 X_i^2$ (iv) Y_i [illegible]

43. Consider the following estimated two variable CLRM [illegible] $0.5X_i$ [illegible] in which n [illegible]

[illegible] $\sum x_i^2 = 2201$, $\sum y_i^2 = 4951$

(i) Obtain the estimated regression coefficient when X is regressed on Y.

(ii) Obtain the coefficient of correlation between X and Y.

(iii) Obtain the unbiased estimate of the error variance when Y is regressed on X.

(iv) Obtain the estimated value of the intercept term and its estimated standard error when Y is regressed on X.

(v) Test the suggestion that Y is positively related to X at 5% level of significance.

44. Consider the following regression equation $Y_i = \alpha + \beta X_i + u_i$, where n = [illegible]

$\sum Y_i = 80$, $\sum x_i^2 = 600$, $\sum y_i^2$ [illegible], $\sum x_i y_i = 480$

(a) Obtain the estimated value of α and β.

(b) Test the hypothesis that X and Y are negatively correlated against the hypothesis that they are not at 5% level of significance.

3

Multiple Linear Regression Model

3.1 Introduction

In simple regression analysis we study the relationship between an explained (dependent) variable Y and an explanatory (independent) variable X. In multiple regression analysis we study the relationship between Y and a number of explanatory variables $X_1, X_2, \dots, X_K$. For example, in demand studies we may be interested in investigating the relationship between quantity demanded of a good and price of that good, prices of substitute goods and income of the consumer. In fact this problem can be analysed with the help of multiple regression analysis.

Let us consider a linear regression model where there are K independent variables say $X_1, X_2, \dots, X_K$ and Y is the only dependent variable. In this case the regression model is given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + u_i \text{, where } i = 1, 2, \dots, n \qquad (3.1)$$

Here $\beta_0, \beta_1, \beta_2, \dots, \beta_K$ are $(K + 1)$ regression parameters

[K = number of explanatory variables and $K + 1$ = number of regression parameters

u = random disturbance term (error term).

β_0 = constant term and $\beta_1, \beta_2, \dots, \beta_K$ are the partial regression coefficients

We make the following assumptions about u_i :

(i) $E(u_i) = 0$ for all i, $i = 1, 2, \dots, n$

(ii) $\text{var}(u_i) = \sigma_u^2$ for all i

(iii) u_i and u_j are independent for all $i \neq j$

(iv) u_i and X are independent for all i and j

(v) u_i is normally distributed for all i [$u_i \sim N(0, \sigma_u^2)$]

(vi) There is no linear dependencies in the explanatory variables.

Under the first four assumptions, we can show that the method of least squares gives estimators of $\beta_0, \beta_1, \beta_2, \dots, \beta_K$ that are unbiased and have minimum variance.

In equation (3.1) if we put $i = 1, 2, \dots, n$, we have

For $i = 1$, $Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \dots + \beta_K X_{K1} + u_1$

„ $i = 2$, $Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \dots + \beta_K X_{K2} + u_2$

„ $i = 3$, $Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \dots + \beta_K X_{K3} + u_3$

… … … … …

„ $i = n$, $Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \dots + \beta_K X_{Kn} + u_n$

[illegible]

$$Y = X\beta + u \qquad (3.2)$$

where

[illegible]

Equation [illegible] represents a set of n equations with k independent explanatory variables with a sample of [illegible].

The model is said to be a Classical Linear Regression Model (CLRM) if it satisfies the following properties:

(i) $E(u)$ is a null vector

$$\text{where } E(u) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0_n$$

since $E(u_i) = 0$ for all $i = 1, 2, \ldots, n$

(ii) The dispersion matrix (variance-covariance matrix) of the disturbance vector is a scalar matrix.

$$D(u) = \sigma_u^2 I_n \text{ where } I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n} = \text{identity matrix and } \sigma_u^2 \text{ is a constant.}$$

Proof Let us suppose u is a random vector variable, then dispersion matrix, $D(u) = E[u - E(u)][u - E(u)]'$

$$= E(uu') \quad \text{as } E(u) = 0$$

$$E\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & u_2 & u_3 & \cdots & u_n \end{bmatrix}$$

$$= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ u_3u_1 & u_3u_2 & u_3^2 \cdots & u_3u_n \\ \vdots & & & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix} = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ E(u_3u_1) & E(u_3u_2) & E(u_3^2) \cdots & E(u_3u_n) \\ \vdots & & & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{bmatrix}$$

$$\begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix}$$

since $E(u_i u_j) = 0$ for $i, j, i \neq j$

and $E(u_i^2) = \sigma_u^2$ [illegible]

$$\text{Hence } [illegible] = \sigma_u^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma_u^2 I_n$$

[illegible] $\sigma_u^2 I_n$

u is a multi variate normal vector with mean 0 and variance covariance matrix $\sigma_u^2 I_n$.

(iv) The independent variables $X_1, X_2, \ldots X_K$ are non stochastic or nonrandom or X is a non-stochastic matrix.

(v) Rank of matrix X is $(K + 1)$.

Rank of a matrix implies the maximum number of linearly independent columns of the matrix.

Since X is a matrix of order $n \times (K + 1)$, all the columns of X should be linearly independent. Now X' is a matrix of order $(K + 1) \times n$ and $(X'X)$ is a matrix of order $(K + 1)(K + 1)$. If the rank of X is $(K + 1)$ then rank of $(X'X)$ is also $(K + 1)$ and $|X'X| \neq 0$ if its rank is $(K + 1)$ and $|X'X| = 0$ if its rank is $< (K + 1)$.

If $|X'X| = 0$ then $(X'X)^{-1}$ does not exist.

§ 2. The Least Squares Method (OLS) for Estimation of Regression Parameters

In vector matrix form the general linear regression model (Equation 3.1) can be written as, $Y = X\beta + u$

$$\text{where } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix}_{n \times (K+1)}$$

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1} \quad \text{and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1) \times 1}$$

Let $\hat{Y} = X\hat{\beta}$ be the vector of the regressed value of Y

$$\text{and } \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}_{(K+1)\times 1} \text{ be the vector of estimators}$$

where $\hat{Y}$ is a $(n \times 1)$ order vector and X is a $n \times (K+1)$ order matrix.

Let e be the residual vector i.e. $e = Y - \hat{Y}$ where $Y = X\beta + u$ and $\hat{Y} = X\hat{\beta}$

$e = Y - X\hat{\beta}$

$$\text{Here, } e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1} \text{ and } e' = [e_1 \; e_2 \; \ldots \; e_n]_{1\times n}$$

$$\text{Now } e'e = [e_1, e_2, \ldots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \quad \therefore e'e = \sum_{i=1}^{n} e_i^2$$

$$e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

Here, $\hat{\beta}'X'Y$ is scalar (1×1), it is equal to its transpose i.e. $\hat{\beta}'X'Y = Y'X\hat{\beta}$

Now by OLS method we have to minimise $\sum_{i=1}^{n} e_i^2 = e'e$ with respect to $\hat{\beta}$

$$\text{Now } \frac{d(e'e)}{d\hat{\beta}} = 0 - X'Y - X'Y + 2X'X\hat{\beta} = 0$$

$$\text{or } 2X'X\hat{\beta} = 2X'Y \quad \text{or} \quad X'X\hat{\beta} = X'Y$$

$$\text{or, } \hat{\beta} = (X'X)^{-1}X'Y, \text{ where } \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}_{(K+1)\times 1}$$

To derive this result more clearly we consider a three variable (with two explanatory variables i.e., when $K = 2$) linear regression model which takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \; i = 1, 2, \ldots, n$$

[illegible]

If we take deviations from respective means, we have

[illegible]

$= \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ [illegible] where [illegible]

and [illegible]

Let us define $X = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix}$

$X' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ and $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

In vector-matrix form the equation,

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ for $i = 1, 2, \ldots, n$, can be written as $Y = X\beta + u$

$$\text{Now } X'X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

$$\text{Now } \beta' X' X \beta = [\beta_1 \ \beta_2] \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$= [\beta_1 \ \beta_2] \begin{bmatrix} \beta_1 \Sigma x_{1i}^2 + \beta_2 \Sigma x_{1i} x_{2i} \\ \beta_1 \Sigma x_{1i} x_{2i} + \beta_2 \Sigma x_{2i}^2 \end{bmatrix}$$

$$= \beta_1^2 \Sigma x_{1i}^2 + \beta_1 \beta_2 \Sigma x_{1i} x_{2i} + \beta_1 \beta_2 \Sigma x_{1i} x_{2i} + \beta_2^2 \Sigma x_{2i}^2$$

$$\beta' X' X \beta = \beta_1^2 \Sigma x_{1i}^2 + 2\beta_1 \beta_2 \Sigma x_{1i} x_{2i} + \beta_2^2 \Sigma x_{2i}^2$$

$$\text{Now } \frac{d}{d\beta}(\beta' X' X \beta) = \begin{bmatrix} \dfrac{d}{d\beta_1}(\beta' X' X \beta) \\ \dfrac{d}{d\beta_2}(\beta' X' X \beta) \end{bmatrix}$$

[illegible]

$$\frac{d}{d\beta}(\beta' X'X\beta) = 2X'X\beta$$

Again $Y'X\beta = (y_1 \; \ldots \; y_n)\begin{bmatrix} x_{11} & x_{21} \\ \vdots & \vdots \\ x_{1n} & x_{2n}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$

$$= (\Sigma x_{1i} y_i \;\; \Sigma x_{2i} y_i)\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = \beta_1 \Sigma x_{1i} y_i + \beta_2 \Sigma x_{2i} y_i$$

Now $\frac{d}{d\beta}(Y'X\beta) = \begin{bmatrix}\frac{d}{d\beta_1}(Y'X\beta)\\ \frac{d}{d\beta_2}(Y'X\beta)\end{bmatrix} = \begin{bmatrix}\Sigma x_{1i}y_i\\ \Sigma x_{2i}y_i\end{bmatrix} = X'Y$

(since $\Sigma x_{1i} y_i$, $\Sigma x_{2i} y_i$ [illegible] $X'Y$)

$$\frac{d}{d\beta}(Y'X\beta) = X'Y$$

Again $\beta'X'Y = (\beta_1 \; \beta_2)\begin{bmatrix}x_{11} & x_{12} & \ldots & x_{1n}\\ x_{21} & x_{22} & \ldots & x_{2n}\end{bmatrix}\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$

$$= (\beta_1 \; \beta_2)\begin{bmatrix}\Sigma x_{1i} y_i\\ \Sigma x_{2i} y_i\end{bmatrix} = \beta_1\Sigma x_{1i}y_i + \beta_2 \Sigma x_{2i} y_i$$

Now $\frac{d}{d\beta}(\beta'X'Y) = \begin{bmatrix}\frac{d}{d\beta_1}(\beta'X'Y)\\ \frac{d}{d\beta_2}(\beta'X'Y)\end{bmatrix} = \begin{bmatrix}\Sigma x_{1i}y_i\\ \Sigma x_{2i}y_i\end{bmatrix} = X'Y$

Thus we have $\frac{d}{d\beta}(\beta'X'X\beta) = 2X'X\beta$

$\frac{d}{d\beta}(Y'X\hat{\beta}) = X'Y$ and $\frac{d}{d\beta}(\beta'X'Y) = X'Y$

Since $e'e = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$

[illegible]

[illegible]

mode

Thus we [illegible] $\hat{\beta}$ [illegible]

$$\text{or}\quad \hat{\beta} = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

For a [illegible] with two explanatory variables [illegible] the equations we can find out the values of $\hat{\beta}_1$ and $\hat{\beta}_2$.

Now $\hat{\beta} = (X'X)^{-1}X'Y$

$$= \frac{1}{|X'X|}\,\text{adj}\,(X'X)\,X'Y$$

Now A_{ij} [illegible] Transpose of matrix [illegible] $(X'X)$

$$= \begin{bmatrix} \Sigma x_{2i}^2 & \Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix} \text{ and } |X'X| = \Sigma x_{1i}^2 \text{ [illegible]}$$

$$\text{Adj}\,(X'X) = \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \frac{1}{|X'X|} \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix} \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{1}{|X'X|}\left[\Sigma x_{2i}^2 \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i}\Sigma x_{2i}y_i\right] = \text{[illegible]}$$

$$\text{and } \hat{\beta}_2 = \frac{1}{|X'X|}\left[-\Sigma x_{1i}x_{2i}\Sigma x_{1i}y_i + \Sigma x_{1i}^2\Sigma x_{2i}y_i\right] = \text{[illegible]}$$

when $\hat{\beta}_1$ and $\hat{\beta}_2$ are known, $\hat{\beta}_0$ can be obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{X}_1 + \hat{\beta}_2\bar{X}_2 \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$$

We can also find out the values of β_1 and β_2 directly by using Cramer's rule

Since $\quad \beta = (X'X)^{-1}X'Y$ or $(X'X)\beta = X'Y$

or [illegible]

or,
$$\begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

or $\hat{\beta}_1 \Sigma x_{1i}^2 + \hat{\beta}_2 \Sigma x_{1i}x_{2i} = \Sigma x_{1i}y_i$ ——— (A)

and $\hat{\beta}_1 \Sigma x_{1i}x_{2i} + \hat{\beta}_2 \Sigma x_{2i}^2 = \Sigma x_{2i}y_i$ ——— (B)

Solving equations (A) and (B) by Cramer's rule we have

$$\hat{\beta}_1 = \frac{\begin{vmatrix} \Sigma x_{1i}y_i & \Sigma x_{1i}x_{2i} \\ \Sigma x_{2i}y_i & \Sigma x_{2i}^2 \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{1i}y_i \, \Sigma x_{2i}^2 - \Sigma x_{1i}x_{2i} \, \Sigma x_{2i}y_i}{\Sigma x_{1i}^2 \, \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

and
$$\hat{\beta}_2 = \frac{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}y_i \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}y_i \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{1i}^2 \, \Sigma x_{2i}y_i - \Sigma x_{1i}x_{2i} \, \Sigma x_{1i}y_i}{\Sigma x_{1i}^2 \, \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ is obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \quad \Rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

Note For the three variable linear regression equation

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$ where $u_i \sim N(0, \sigma_u^2)$, we can also find out the values of the regression parameters in another way. This method is described below.

The estimated regression line is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \text{ and } \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

where $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS estimates of β_0, β_1 and β_2.

We obtain by subtraction $\hat{y}_i = \hat{Y}_i - \bar{Y}$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

or $\hat{Y}_i = \hat\beta_1 x_{1i} + \hat\beta_2 x_{2i} + \bar{Y}$

or $\hat{y}_i = \hat\beta_1 x_{1i} + \hat\beta_2 x_{2i}$ where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$, $y_i = Y_i - \bar{Y}$

Now errors of estimate $e_i = y_i - \hat{y}_i = y_i - \hat\beta_1 x_{1i} - \hat\beta_2 x_{2i}$

and $\sum e_i^2 = \sum (y_i - \hat\beta_1 x_{1i} - \hat\beta_2 x_{2i})^2$

The first order conditions for minimization require

$$\frac{\partial \sum e_i^2}{\partial \hat\beta_1} = -2\sum (y_i - \hat\beta_1 x_{1i} - \hat\beta_2 x_{2i})(x_{1i}) = 0$$

or $\sum x_{1i} y_i = \hat\beta_1 \sum x_{1i}^2 + \hat\beta_2 \sum x_{1i} x_{2i}$ ...(1)

and $$\frac{\partial \sum e_i^2}{\partial \hat\beta_2} = -2\sum (y_i - \hat\beta_1 x_{1i} - \hat\beta_2 x_{2i})(x_{2i}) = 0$$

or $\sum x_{2i} y_i = \hat\beta_1 \sum x_{1i} x_{2i} + \hat\beta_2 \sum x_{2i}^2$ ...(2)

Now solving equations (1) and (2) by Cramer's rule we have,

$$\hat\beta_1 = \frac{\begin{vmatrix} \sum x_{1i} y_i & \sum x_{1i} x_{2i} \\ \sum x_{2i} y_i & \sum x_{2i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{vmatrix}} = \frac{\sum x_{1i} y_i \sum x_{2i}^2 - \sum x_{2i} y_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

and $$\hat\beta_2 = \frac{\begin{vmatrix} \sum x_{1i}^2 & \sum x_{1i} y_i \\ \sum x_{1i} x_{2i} & \sum x_{2i} y_i \end{vmatrix}}{\begin{vmatrix} \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{vmatrix}} = \frac{\sum x_{1i}^2 \sum x_{2i} y_i - \sum x_{1i} y_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

When $\hat\beta_1$ and $\hat\beta_2$ are known, $\hat\beta_0$ can be obtained from the relation

$$\bar{Y} = \hat\beta_0 + \hat\beta_1 \bar{X}_1 + \hat\beta_2 \bar{X}_2 \quad \text{i.e.} \quad \hat\beta_0 = \bar{Y} - \hat\beta_1 \bar{X}_1 - \hat\beta_2 \bar{X}_2$$

Example 3.1 Consider the following regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where u_i is normally distributed with mean 0 and variance σ_u^2.

Y	4	7	3	9	7
X_1	2	3	1	4	9
X_2	5	3	2	6	7

Estimate $\hat\beta_0$, $\hat\beta_1$ and $\hat\beta_2$ (the OLS estimators of β_0, β_1 and β_2).

Solution Calculations for the estimators of the regression parameters

Table 3.1

[illegible]

Here n = 5 as five sets of values of the variables are given.

Now [illegible]

and [illegible]

We know that the values of β_1 and β_2 in the regression equation

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$ can be obtained by the OLS method. The estimators of β_1 and β_2 are $\hat{\beta}_1$ and $\hat{\beta}_2$,

where $\hat{\beta}_1 =$ [illegible]

[illegible] $= \frac{5[illegible]}{619} =$ [illegible]

and $\hat{\beta}_2 =$ [illegible]

[illegible] $= 0.720$

$\hat{\beta}_1 = 0.806$ and $\hat{\beta}_2 = 0.720$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ can be obtained from the relation

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ [illegible] $= 5.368$

$\hat{\beta}_0 = 5.368$, $\hat{\beta}_1 = 0.806$ and $\hat{\beta}_2 = 0.720$

3.2.2 The Regression Coefficients Expressed in terms of Variances (SDs) and Coefficient of Correlations

In a three variable linear regression model ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$) [illegible]

[illegible]

Similarly [illegible]

and [illegible]

Now putting values in the expression of β_1 we get

$$\beta_1 = [illegible]$$

$$= [illegible]$$

$$= [illegible]$$

$$= [illegible]$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\left[[illegible] \right]$$

Proceeding in the same way we get [illegible]

[illegible] where the symbols have the usual meaning.

3.2.2 Determination of Variances and Covariances of the Estimators of the Regression Parameters in Three Variable Linear Regression Model

For a three variable linear regression model, we assume the regression equation of the form

$$= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad [illegible]$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{u} \quad \text{where } \bar{u} = 0$$

Now [illegible]

or, $= \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ where $x_{1i} = X_{1i} - \bar{X}_1$, [illegible] and $u_i =$ [illegible]

In vector-matrix form the set of n equations [illegible] can be written as $Y = X\beta + u$

where $X =$ [illegible] and $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

Now $X'X =$ [illegible]

Now [illegible]

where $E(\hat{\beta}_1) = \beta_1$ and [illegible]

$$\begin{bmatrix} \mathrm{var}(\hat{\beta}_1) & \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) & \mathrm{var}(\hat{\beta}_2) \end{bmatrix} = \sigma_u^2 (X'X)^{-1}$$

[See Property [illegible] of OLS estimator vector $\hat{\beta}$]

$$= \sigma_u^2 \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}^{-1}$$

$$= \frac{\sigma_u^2}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} \mathrm{Adj} \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}^{-1}$$

Now, $\begin{bmatrix} \mathrm{var}(\hat\beta_1) & \mathrm{cov}(\hat\beta_1\hat\beta_2) \\ \mathrm{cov}(\hat\beta_1\hat\beta_2) & \mathrm{var}(\hat\beta_2) \end{bmatrix} = \dfrac{\sigma_u^2}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2}\begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix}$

$$= \sigma_u^2 \begin{bmatrix} \dfrac{\Sigma x_{2i}^2}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2} & \dfrac{-\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2} \\ \dfrac{-\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2} & \dfrac{\Sigma x_{1i}^2}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2} \end{bmatrix}$$

$$\mathrm{var}(\hat\beta_1) = \frac{\sigma_u^2\Sigma x_{2i}^2}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2} = \frac{\sigma_u^2 n\sigma_{x_2}^2}{n\sigma_{x_1}^2 n\sigma_{x_2}^2 - n^2 r_{x_1x_2}^2\sigma_{x_1}^2\sigma_{x_2}^2}$$

$$\mathrm{var}(\hat\beta_1) = \frac{\sigma_u^2 n\sigma_{x_2}^2}{n\sigma_{x_1}^2 n\sigma_{x_2}^2(1-r_{x_1x_2}^2)} = \frac{\sigma_u^2}{n\sigma_{x_1}^2(1-r_{x_1x_2}^2)}$$

[Since $\Sigma x_{1i}^2 = \Sigma(X_{1i}-\bar X_1)^2 = n\frac{1}{n}\Sigma(X_{1i}-\bar X_1)^2 = n\sigma_{x_1}^2$, similarly, $\Sigma x_{2i}^2 = n\sigma_{x_2}^2$

and $\Sigma x_{1i}x_{2i} = \Sigma(X_{1i}-\bar X_1)(X_{2i}-\bar X_2) = n\frac{1}{n}\Sigma(X_{1i}-\bar X_1)(X_{2i}-\bar X_2)$

$= n\,\mathrm{cov}(X_1, X_2) = n\,r_{x_1x_2}\sigma_{x_1}\sigma_{x_2}$ as $\dfrac{\mathrm{cov}(X_1, X_2)}{\sigma_{x_1}\sigma_{x_2}} = r_{x_1x_2}$.]

Similarly, $\mathrm{var}(\hat\beta_2) = \dfrac{\sigma_u^2\Sigma x_{1i}^2}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2}$

$$= \frac{\sigma_u^2 n\sigma_{x_1}^2}{n\sigma_{x_1}^2 n\sigma_{x_2}^2 - n^2 r_{x_1x_2}^2\sigma_{x_1}^2\sigma_{x_2}^2} = \frac{\sigma_u^2 n\sigma_{x_1}^2}{n\sigma_{x_1}^2 n\sigma_{x_2}^2(1-r_{x_1x_2}^2)}$$

$$\mathrm{var}(\hat\beta_2) = \frac{\sigma_u^2}{n\sigma_{x_2}^2(1-r_{x_1x_2}^2)}$$

Again, $\mathrm{cov}(\hat\beta_1\hat\beta_2) = \dfrac{-\sigma_u^2\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2}$

$$= \frac{-\sigma_u^2 n\,r_{x_1x_2}\sigma_{x_1}\sigma_{x_2}}{n\sigma_{x_1}^2 n\sigma_{x_2}^2 - n^2 r_{x_1x_2}^2\sigma_{x_1}^2\sigma_{x_2}^2} = \frac{-\sigma_u^2 r_{x_1x_2}\sigma_{x_1}\sigma_{x_2}}{n\sigma_{x_1}^2\sigma_{x_2}^2(1-r_{x_1x_2}^2)}$$

$\text{cov}(\hat\beta_1, \hat\beta_2) =$ [illegible]

Now, $\text{var}(\hat\beta_1 + \hat\beta_2) = \text{var}(\hat\beta_1) + \text{var}(\hat\beta_2) + 2\,\text{cov}(\hat\beta_1, \hat\beta_2)$

$$=$$ [illegible]

and $\text{var}(\hat\beta_1 - \hat\beta_2) = \text{var}(\hat\beta_1) + \text{var}(\hat\beta_2) - 2\,\text{cov}(\hat\beta_1, \hat\beta_2)$

$$=$$ [illegible]

since $\bar{Y} = \hat\beta_0 + \hat\beta_1 \bar{X}_1 + \hat\beta_2 \bar{X}_2 \Rightarrow \hat\beta_0 = \bar{Y} - \hat\beta_1 \bar{X}_1 - \hat\beta_2 \bar{X}_2$

It can be seen that

$$\text{var}(\hat\beta_0) = \frac{\sigma_u^2}{n} + \bar{X}_1^2\,\text{var}(\hat\beta_1) + 2\bar{X}_1\bar{X}_2\,\text{cov}(\hat\beta_1, \hat\beta_2) + \bar{X}_2^2\,\text{var}(\hat\beta_2)$$

$$\text{cov}(\hat\beta_0, \hat\beta_1) = -[\bar{X}_1\,\text{var}(\hat\beta_1) + \bar{X}_2\,\text{cov}(\hat\beta_1, \hat\beta_2)]$$

and $\text{cov}(\hat\beta_0, \hat\beta_2) = -[\bar{X}_1\,\text{cov}(\hat\beta_1, \hat\beta_2) + \bar{X}_2\,\text{var}(\hat\beta_2)]$

Note. In calculating $\text{var}(\hat\beta_0)$, $\text{var}(\hat\beta_1)$, $\text{var}(\hat\beta_2)$, $\text{cov}(\hat\beta_1, \hat\beta_2)$, $\text{cov}(\hat\beta_0, \hat\beta_1)$ and $\text{cov}(\hat\beta_0, \hat\beta_2)$, if σ_u^2 is not known it is to be replaced by its unbiased estimator $\hat\sigma_u^2 = \Sigma e_i^2/(n-3)$.

Example 3.2. The following table presents data on a sample of five persons randomly drawn from a large firm giving their annual salaries in thousands of dollars (Y), years of education (X_1) and years of experience with the firm they are working (X_2).

Y	30	20	36	24	40
X_1	4	3	6	4	5
X_2	10	8	11	9	12

Assuming a linear regression of the form

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $u_i \sim N(0, \sigma_u^2)$

(i) find the OLS estimators $\hat\beta_0$, $\hat\beta_1$ and $\hat\beta_2$

(ii) find the value of r_{12}

(iii) find the estimated regression equation

(iv) find Σe_i^2

(v) find the values of $\text{var}(\hat\beta_1)$, $\text{var}(\hat\beta_2)$ and $\text{var}(\hat\beta_0)$

(vi) find the value of $\text{cov}(\hat\beta_1, \hat\beta_2)$

Solution :

Calculation Table 3.2

Y	X_{1i}	X_{2i}	$y_i = Y_i - \bar{Y}$	y_i^2	$x_{1i} = X_{1i} - \bar{X}_1$	x_{1i}^2	$x_{2i} = X_{2i} - \bar{X}_2$	x_{2i}^2	$x_{1i}y_i$	$x_{2i}y_i$	[illegible]	[illegible]	[illegible]
30	4	10	0	0	1	1	0	0	0	0	[illegible]	[illegible]	[illegible]
20	3	8	-10	100	2	4	2	4	20	20	4	[illegible]	[illegible]
36	6	11	6	36	1	1	1	1	6	6	[illegible]	[illegible]	[illegible]
24	4	9	-6	36	1	1	1	1	6	6	[illegible]	[illegible]	[illegible]
40	8	12	10	100	3	9	2	4	30	20	6	[illegible]	[illegible]
$\Sigma Y_i = 150$	$\Sigma X_{1i} = 25$	$\Sigma X_{2i} = 50$	$\Sigma y_i = 0$	$\Sigma y_i^2 = 272$	$\Sigma x_{1i} = 0$	$\Sigma x_{1i}^2 = 16$	$\Sigma x_{2i} = 0$	$\Sigma x_{2i}^2 = 10$	$\Sigma x_{1i}y_i = 62$	$\Sigma x_{2i}y_i = 52$	[illegible]	[illegible]	[illegible]

Here $n = 5$ as five sets of values are given.

Now $\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{150}{5} = 30$, $\bar{X}_1 = \frac{\Sigma X_{1i}}{n} = \frac{25}{5} = 5$, $\bar{X}_2 = \frac{\Sigma X_{2i}}{n} = \frac{50}{5} = 10$

We have to find out the OLS estimators β_0, β_1 and β_2.

We know that $\beta_1 =$ [illegible]

We now put the values from the calculation table and get

$\beta_1 =$ [illegible] $= \frac{-4}{16} = -0.25$

Similarly $\beta_2 =$ [illegible] $= \frac{88}{16} = 5.5$

When β_1 and β_2 are known, β_0 can be obtained from the relation

$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 = 30 -$ [illegible]

$\beta_0 = 30 +$ [illegible] $= 23.75$

Thus the OLS estimators of the parameters are $\beta_0 = 23.75$, $\beta_1 = -0.25$ and $\beta_2 = 5.5$

(ii) We have to find out the value of product moment correlation coefficient between two explanatory variables X_1 and X_2 i.e. $r_{X_1X_2}$

We know that $r_{X_1X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} =$ [illegible]

$=$ [illegible] $= \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2}\sqrt{\Sigma x_2^2}} =$ [illegible] $= 0.45$ $\quad r_{X_1X_2} = 0.45$

(iii) The estimated regression line is given by

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$

$\hat{Y}_i = 23.75 - 0.25 X_{1i} + 5.5 X_{2i}$ is the estimated regression line equation

(iv) We have to find out the value of Σe_i^2

where $\Sigma e_i^2 = \Sigma (Y_i - \hat{Y}_i)^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + (Y_3 - \hat{Y}_3)^2$

$+ (Y_4 - \hat{Y}_4)^2 + (Y_5 - \hat{Y}_5)^2$

$= (-0.25)^2 + (0.5)^2 + (0.75)^2 + (-0.75)^2 +$ [illegible]

In particular when $Y = Y_1 = 30$, $X_{11} = 4$, $X_{21} = 10$ then

$\hat{Y}_i = \hat{Y}_1 = 23.75 - 0.25 \times 4 + 5.5 \times 10 =$ [illegible]

[illegible]

When $Y = Y_3 =$ [illegible]

then $\hat{Y} = \hat{Y}_3 =$ [illegible] $= 23.75$ [illegible]

$e_3 = Y_3 - \hat{Y}_3 =$ [illegible]

When $Y = Y_4 = 24$, [illegible]

then $\hat{Y} = \hat{Y}_4 =$ [illegible]

$e_4 = Y_4 - \hat{Y}_4 = 24 -$ [illegible]

When $Y = Y_5 = 40$, [illegible]

then $\hat{Y} = \hat{Y}_5 =$ [illegible]

$= 23.75$ [illegible]

$e_5 = Y_5 - \hat{Y}_5 = 40 - 40.75 = -0.75$

v) Now we have to calculate the variances of the OLS estimators of the regression parameters, $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$ and $var(\hat{\beta}_3)$.

We know that $var(\hat{\beta}_1) = \frac{\sigma_u^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$ [illegible] Here σ_u^2 is not known and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 =$ [illegible]

Here $\frac{\sum e_i^2}{n-3} = \frac{1.5}{5-3} = \frac{1.5}{2} = 0.75$

$$var(\hat{\beta}_1) = \frac{\left(\frac{\sum e_i^2}{n-3}\right) \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} = \frac{0.75 \times 10}{16 \times \text{[illegible]}}$$

$= \frac{7.5}{16} = 0.4687$ $var(\hat{\beta}_1) = 0.4687$

Similarly, var($\hat{\beta}_2$) = [illegible] 0.75

var($\hat{\beta}_2$) = 0.75

and var($\hat{\beta}_1$) [illegible] var [illegible] cov($\hat{\beta}_1$, $\hat{\beta}_2$) [illegible] var [illegible]

We find $\hat{\sigma}_u^2$ = [illegible] = 0.75 [illegible]

var($\hat{\beta}$) = 0.0667, var($\hat{\beta}$) = 0.75 and cov($\hat{\beta}$, $\hat{\beta}$) = 0.5625 obtained from [illegible]

Thus we have var [illegible] 0.0667 [illegible] 0.5625 [illegible]

[illegible] 0.75 [illegible] 0.75

var($\hat{\beta}_0$) = [illegible]

(iii) We have to find out the value of cov($\hat{\beta}_1$, $\hat{\beta}_2$)

Now cov($\hat{\beta}_1$, $\hat{\beta}_2$) = [illegible] Here σ_u^2 is not known and hence replaced by its unbiased estimator $\hat{\sigma}_u^2 = \frac{\sum e^2}{n-k} = 0.75$

cov($\hat{\beta}_1$, $\hat{\beta}_2$) = [illegible] $= \frac{9}{16}$ = 0.5625

cov($\hat{\beta}_1$, $\hat{\beta}_2$) = 0.5625

3.3 Properties of OLS Estimator Vector $\hat{\beta}$

Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$ be [illegible] Multiple Linear Regression model

In vector-matrix form the model takes the form $Y = X\beta + u$

$$\text{where } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & & & & \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n \times k}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{K \times 1} \quad \text{and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1}$$

[illegible]

[illegible]

Property 3 : $\hat{\beta}$ is an unbiased estimator of β

Proof: [illegible]

[illegible]

[illegible]

where [illegible] matrix X as well as $X'X$

[illegible]

Again $E(\hat{\beta}) = E[\beta + (X'X)^{-1}X'u]$

$$= \beta + (X'X)^{-1}X'E(u) = E(\beta) \quad [\because E(u) = 0]$$

$E(\hat{\beta}) = \beta$ where $E(u) = E$ [illegible]

This shows that the OLS estimator of β is an unbiased estimator [illegible]

$$\text{Here } E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ \vdots \\ E(\hat{\beta}_K) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \Rightarrow \begin{matrix} E(\hat{\beta}_0) = \beta_0 \\ E(\hat{\beta}_1) = \beta_1 \\ E(\hat{\beta}_2) = \beta_2 \\ \vdots \\ E(\hat{\beta}_K) = \beta_K \end{matrix}$$

This implies that OLS estimator of each parameter is an unbiased estimator. In terms of a linear regression model with two explanatory variables we have $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad i = 1, 2, \ldots, n$

$$\text{In this model we have } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{3\times 1}$$

$$\text{and } E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}$$

$$E(\hat{\beta}) = \beta \text{ where } E(\hat{\beta}_0) = \beta_0, \; E(\hat{\beta}_1) = \beta_1, \; E(\hat{\beta}_2) = \beta_2$$

Property 2. The Dispersion matrix or variance-covariance matrix of $\hat{\beta}$ is given by $\sigma_u^2(X'X)^{-1}$.

Proof. By definition, dispersion matrix or variance covariance matrix of $\hat{\beta}$ is given by $D(\hat{\beta}) = E[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'$ where $E(\hat{\beta}) = \beta$

$$= \begin{bmatrix} E(\hat{\beta}_0 - \beta_0)^2 & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0) & \cdots & E(\hat{\beta}_K - \beta_K)(\hat{\beta}_0 - \beta_0) \\ E(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) & E(\hat{\beta}_1 - \beta_1)^2 & \cdots & E(\hat{\beta}_K - \beta_K)(\hat{\beta}_1 - \beta_1) \\ \vdots & \vdots & & \vdots \\ E(\hat{\beta}_0 - \beta_0)(\hat{\beta}_K - \beta_K) & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_K - \beta_K) & \cdots & E(\hat{\beta}_K - \beta_K)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \cdots & \text{cov}(\hat{\beta}_K, \hat{\beta}_0) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_K, \hat{\beta}_1) \\ \vdots & \vdots & & \vdots \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_K) & \text{cov}(\hat{\beta}_1, \hat{\beta}_K) & \cdots & \text{var}(\hat{\beta}_K) \end{bmatrix}$$

Here the diagonal terms are variances and non-diagonal terms are covariances. So, it is also called variance covariance matrix.

$$D(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

$$= E[\{(X'X)^{-1}X'u\}\{(X'X)^{-1}X'u\}']$$

Since $\hat{\beta} = \beta + (X'X)^{-1}X'u$

$\hat{\beta} - \beta = (X'X)^{-1}X'u$

$$= E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$$

$$= (X'X)^{-1}X'E(uu')X(X'X)^{-1}$$

$$= (X'X)^{-1}X'\sigma_u^2 I_n X(X'X)^{-1}$$

$$= \sigma_u^2(X'X)^{-1}X'X(X'X)^{-1}$$

$$= \sigma_u^2(X'X)^{-1}I_{K+1} = \sigma_u^2(X'X)^{-1}$$

$$D(\hat{\beta}) = \sigma_u^2(X'X)^{-1}$$

Since $D(u)$

$= E[u - E(u)][u - E(u)]'$

$= E(uu')$ as $E(u) = 0$

$= \sigma_u^2 I_n$ (See 3.1 (ii)),

where I_{K+1} = Identity matrix of order $(K+1) \times (K+1)$

Proceeding in the same way we can also derive the result $D(\hat{\beta}) = \sigma_u^2(X'X)^{-1}$ for a regression model with two explanatory variables

i.e., $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$; $i = 1, 2, 3, \ldots, n$

Property 3. *j*th element of $\hat{\beta}$ is the best linear unbiased estimator of the *j*th element of β. Alternatively, $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) of β.

Proof. Since, $\hat{\beta} = (X'X)^{-1}X'Y$

[illegible]

[illegible]

[illegible]

[illegible] β_0 [illegible]

Let us define $\beta_0^* = C'Y$ where [illegible]

Now we have to find out the conditions under which β_0^* is an unbiased estimator of β_0.

Now, $\beta_0^* = C'Y = C'(X\beta + u)$ since $Y = X\beta + u$

$$= C'X\beta + C'u$$

$$E(\beta_0^*) = E[C'X\beta] + E(C'u)$$

$$= C'X\beta + 0 = C'X\beta \quad [\because E(u) = 0_{n\times 1}]$$

$$E(\beta_0^*) = C'X\beta$$

Now $E(\beta_0^*) = \beta_0$ if $C'X = e_1'$

$$\therefore e_1'\beta = [1 \ 0 \ \cdots \ 0]\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_K\end{bmatrix} = \beta_0$$

This shows that β_0^* is an unbiased estimator of β_0. The condition for β_0^* to be an unbiased estimator of β_0 is given by $C'X = e_1$

Again, $\beta_0^* = C'X\beta + C'u$

$$= e_1'\beta + C'u = \beta_0 + [C_1 \ C_2 \ \cdots \ C_n]\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{bmatrix}$$

or, $\beta_0^* - \beta_0$ $\sum$ [illegible] β_0^* β_0 $\sum$ [illegible]

Now var $(\beta_0^*) = E[\beta_0^*$ [illegible] $E(\beta_0^*) = \beta_0$

$$E\left(\sum c_i u_i\right) = \sigma_u \sum c_i \quad [illegible]$$

var $\beta_0^* = \sigma_u \sum c_i$

Now we have to minimise var (β_0^*) subject to the condition that [illegible] through the choice of the vector C. In other words we have to minimise $\sigma \sum$ [illegible] subject to the condition $C'X = e_1'$

For the sake of simplicity we put $\sigma_u^2 = \frac{1}{2}$

The Lagrangean is given by,

$$L = \frac{1}{2}\sum_{i=1}^{n} C_i^2 - (C'X - e_1')\lambda$$

where $\lambda = \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix}_{(K+1)\times 1}$ is the vector of Lagrangean multipliers

$$L = \frac{1}{2}C'C - [C'X - e_1']\lambda$$

Now differentiating it with respect to C we get

$\frac{\partial L}{\partial C} = C - X\lambda = 0_{n\times 1}$, a null column vector where [illegible]

$\frac{\partial}{\partial C}[C'C] = 2C$ and $\frac{\partial L}{\partial \lambda} = C'X - e_1' = 0_{1\times(K+1)}$, a null row vector [illegible]

$$\frac{\partial L}{\partial C} = C - X\lambda = 0_{n\times 1}$$

or $C = X\lambda$ i.e. $C' = \lambda'X'$ and $C'X = \lambda'X'X$

Again $\frac{\partial L}{\partial \lambda} = C'X - e_1' = 0_{1\times(K+1)}$ $\therefore e_1' = C'X$

$e_1' = C'X = \lambda'X'X$ or, $\lambda'X'X = e_1'$

or, $\lambda' = e_1'(X'X)^{-1}$ $\therefore C' = \lambda'X' = e_1'(X'X)^{-1}X'$

So β_0^* [illegible]

or β_1^* [illegible] β_0 [illegible] $\begin{bmatrix}\beta_0\\\beta\\\vdots\\\beta_k\end{bmatrix}$

So, it is proved that under the condition that β^* is an unbiased [illegible] variance of β_1^* i.e. var (β_1^*) is minimum when $\beta_1^* = \beta_1$.

Applying the similar mathematical technique it can be proved that

var (β_1^*) is minimum when $\beta_1^* = \beta_1$

var (β_2^*) is minimum when $\beta_2^* = \beta_2$

var (β_k^*) is minimum when $\beta_k^* = \beta_k$

Hence it is proved that OLS estimators are the best linear unbiased estimators of the regression parameters i.e. $\hat{\beta}$ is the BLUE of β. This is known as the GAUSS-MARKOV THEOREM.

In case of three variable linear regression model ($Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, $i = 1, 2, \ldots, n$) we can prove that $\hat{\beta}$ is the BLUE of β where

$$\beta = \begin{bmatrix}\beta_1\\\beta_2\\\beta_3\end{bmatrix} \text{ and } \hat{\beta} = \begin{bmatrix}\hat{\beta}_1\\\hat{\beta}_2\\\hat{\beta}_3\end{bmatrix}$$

Property 4 Unbiased estimator of σ_u^2 is $\frac{e'e}{n-(k+1)} = \sum_{i=1}^{n} e_i^2 / [n-(k+1)]$ **where k = number of explanatory variables and $(k+1)$ = number of parameters including the constant intercept term.**

Proof Since $Y = X\beta + u$, $\hat{Y} = X\hat{\beta}$ and $Y = \hat{Y} + e$

or, $e = Y - \hat{Y}$ or $e = Y - X\hat{\beta}$ $e' = Y' - \hat{\beta}'X'$

Now $e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$

$= [Y - X(X'X)^{-1}X'Y]'[Y - X(X'X)^{-1}X'Y]$ since $\hat{\beta} = (X'X)^{-1}X'Y$

$= Y'[I - X(X'X)^{-1}X']'[I - X(X'X)^{-1}X']Y$

$= Y'[I - X(X'X)^{-1}X'][I - X(X'X)^{-1}X']Y$

$= Y'MMY$ where $M = I - X(X'X)^{-1}X'$

$e'e = Y'M'MY = Y'MY$ | I = Identity matrix

$= (X\beta + u)'M(X\beta + u)$ | where M is idempotent matrix

$= (\beta'X' + u')M(X\beta + u)$ | for which $M^2 = M$

$= \beta'X'MX\beta + u'MX\beta + \beta'X'Mu + u'Mu$

Now if we put $M = I - X(X'X)^{-1}X'$ then

$\beta'X'MX\beta + u'MX\beta + \beta'X'Mu = 0$

And hence $e'e = u'Mu = u'[I - X(X'X)^{-1}X']u$

$$= u'u - u'X(X'X)^{-1}X'u = \sum_{i=1}^{n} u_i^2 - u'X(X'X)^{-1}X'u$$

$E(e'e) = \Sigma E(u_i^2) - E[u'u]\,\text{trace}\,[X(X'X)^{-1}X']$ $[E(u_i^2) = \sigma_u^2$ and $u'u = \sum_{i=1}^{n} u_i^2]$

$= n\sigma_u^2 - \sigma_u^2(K+1)$ [Here trace $X(X'X)^{-1}X' = K+1$]

$E(e'e) = \sigma_u^2[n - (K+1)]$

or, $E\left[\frac{e'e}{n-(K+1)}\right] = \sigma_u^2$ or, $E\left[\frac{\Sigma e_i^2}{n-(K+1)}\right] = \sigma_u^2$

This proves that $\frac{e'e}{n-(K+1)}$ or $\frac{\Sigma e_i^2}{n-(K+1)}$ is the unbiased estimator of σ_u^2.

In particular for a linear regression model with two explanatory variables

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

we have $E\left[\frac{\Sigma e_i^2}{n-3}\right] = \sigma_u^2$ or, $E(\hat{\sigma}_u^2) = \sigma_u^2$ where $\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3}$ and $K = 2$

3.4. MLE of β and σ_u^2 in the Multiple Regression Model

Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i$ for $i = 1, 2, \dots, n$ be the equation of the general linear regression model. In vector matrix form the set of n equations can be written as, $Y = X\beta + u$

where $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$ $X = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{K1} \\ 1 & X_{12} & X_{22} & X_{K2} \\ \vdots & & & \\ 1 & X_{1n} & X_{2n} & X_{Kn} \end{bmatrix}_{n\times K}$

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1}$ $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}$

[illegible]

[illegible] We also assume that [illegible] $u_1, u_2, \ldots, u_n$ are independent.

[illegible] is the joint probability density functions of the [illegible] u_n and [illegible] the p.d.f. of u_i for $i = 1, 2, \ldots, n$

Since $u_1, u_2, \ldots, u_n$ are independent [illegible] $= \prod f_i(u_i)$

Since u_i [illegible] σ_u then

$$f(u_i) = \frac{1}{\sqrt{2\pi}\,\sigma_u} e^{[illegible]}$$

$$\text{So, } f(u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_u} e^{[illegible]}$$

This is the likelihood function of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ_u and is denoted by

$$L(\beta', \sigma_u) = \frac{1}{(\sqrt{2\pi})^n \sigma_u^n} e^{-\frac{\sum_{i=1}^{n} u_i^2}{2\sigma_u^2}} \quad \text{where } \beta' = [\beta_0, \beta_1, \ldots, \beta_k]_{1\times(k+1)}$$

$$\text{Now } \log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2$$

Now to obtain the MLE of the parameters Log L is to be maximised with respect to the parameters.

(a) To obtain the MLE of β we have to maximise Log L with respect to β which is equivalent to minimization of $\sum_{i=1}^{n} u_i^2$ with respect to β.

So, we have to minimise Σu_i^2 through the choice of β

$$\text{Since } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} \qquad u' = [u_1\ u_2 \ldots u_n]_{1\times n}$$

$$u'u = \sum_{i=1}^{n} u_i^2$$

We know that [illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible] the [illegible]

[illegible]

[illegible] identical with the [illegible] means that MLE of β equals OLS of β [illegible]

[illegible] the MLE of σ_u

In order to derive the MLE of σ_u we have to maximize

$$\log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum u_i^2 \quad \text{with respect to } \sigma_u$$

$$\text{Now } \log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2$$

$$= -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}(Y - X\beta)'(Y - X\beta)$$

Since [illegible]

and [illegible]

Since it is proved that MLE of β [illegible] = OLS of β

$$\log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}(Y - X\beta)'(Y - X\beta)$$

$$= -\frac{n}{2}\log 2\pi - n\log\sigma_u - [illegible]$$

(Since [illegible] and [illegible] and $e'e = \sum e_i^2$

[illegible]

or $\frac{n}{\sigma_u^2}$ [illegible]

So $\sum_{i=1}^{n} e_i^2 / n$ is the MLE of σ_u^2

It should be noted that MLE of σ_u^2 is not an unbiased estimator of σ_u^2 but a consistent or asymptotically unbiased estimator of σ_u^2

Since MLE of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 / n$

$$\text{Now} \quad \sum_{i=1}^{n} e_i^2 / n = \frac{n-(K+1)}{n} \cdot \frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)} = \left(1-\frac{K+1}{n}\right) \cdot \frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$$

$$\text{or} \quad \sum_{i=1}^{n} e_i^2 / n = \left(1-\frac{K+1}{n}\right) \frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$$

As $n \to \infty$, $\frac{(K+1)}{n} \to 0$ and hence

$\sum_{i=1}^{n} e_i^2 / n \to \sum_{i=1}^{n} e_i^2 / (n-(K+1))$ where $\frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$ is an unbiased estimator of σ_u^2

This proves that MLE of σ_u^2 i.e. $\sum_{1}^{n} e_i^2 / n$ is an asymptotically unbiased or consistent estimator of σ_u^2. In a three variable (with two explanatory variables, i.e., when $K = 2$) linear regression model we have MLE of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 / n$ and unbiased estimator of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 / (n-3)$

3.5 Expression of Multiple Correlation Coefficient in the General Linear Regression Model

Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$ be the regression [illegible] where $i = 1, 2, \dots, n$.

Here Y is regressed on $X_1, X_2, \dots, X_k$. So the multiple correlation coeff. [illegible] denoted by the symbol,

$$R_{Y\cdot X_1 X_2 \dots X_k} = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = \frac{\sum(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum(Y_i - \bar{Y})^2} \qquad \text{Since } \bar{\hat{Y}} = \bar{Y}$$

$$= \frac{\sum \hat{Y}_i^2 - n\bar{Y}^2}{\sum Y_i^2 - n\bar{Y}^2} = \frac{\hat{Y}'\hat{Y} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

$$= \frac{\hat{\beta}' X' X \hat{\beta} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} \qquad \text{where } \hat{Y} = X\hat{\beta} \text{ and } [illegible]$$

$$= \frac{\hat{\beta}'(X'X\hat{\beta}) - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} = \frac{\hat{\beta}' X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

where $\hat{\beta} = (X'X)^{-1} X'Y$ [illegible]

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$X'X\hat{\beta} = X'Y$$

3.6 The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation in the Three-Variable Linear Regression Model

In the two-variable case we have seen that R^2 (or r^2) measures the goodness of fit of the regression equation ($Y_i = \alpha + \beta X_i + u_i$, $i = 1, 2, \dots, n$); that is, it gives the proportion or percentage of the total variation in the dependent variable Y explained by the single explanatory variable X. This notation of R^2 can be easily extended to regression models containing more than two variables. Thus, in the three-variable model we would like to know the proportion of the variation in Y explained by the variables X_1 and X_2 jointly.

This quantity that gives this information is known as the *multiple coefficient of determination* and is denoted by $R^2_{Y\cdot X_1 X_2}$ or simply R^2 and conceptually it is similar to r^2.

The estimated three-variable regression line ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \dots, n$) is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$, where $\hat{Y}_i$ is the estimated value of Y from the fitted regression line and is an estimator of true $E(Y_i \mid X_{1i}, X_{2i})$ and $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$.

Taking deviations from means we have

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

or $Y_i - \bar{Y} = \hat{\beta}_1(X_{1i} - \bar{X}_1) + \hat{\beta}_2(X_{2i} - \bar{X}_2) + e_i$

or $y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$

where $y_i = Y_i - \bar{Y}$, $x_{1i} = X_{1i} - \bar{X}_1$ and $x_{2i} = X_{2i} - \bar{X}_2$

Now errors of estimate $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$

$y_i = \hat{y}_i + e_i$

Now $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 + 2\Sigma \hat{y}_i e_i$

or $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$ as $\Sigma \hat{y}_i e_i = 0$

i.e TSS = ESS + RSS

when TSS = Total Sum of Squares

ESS = Explained Sum of Squares

RSS = Residual Sum of Squares

Now by definition, $R^2_{Y \cdot X_1 X_2} = R^2 = \dfrac{\Sigma \hat{y}_i^2}{\Sigma y_i^2} = \dfrac{ESS}{TSS}$

$$= \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} = \frac{\Sigma y_i^2 - \Sigma e_i^2}{\Sigma y_i^2}$$

$$= 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} = 1 - \frac{RSS}{TSS}$$

Since

$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$

$\Sigma \hat{y}_i^2 = \Sigma y_i^2 - \Sigma e_i^2$

Since $e_i = y_i - \hat{y}_i$, $\Sigma e_i^2 = \Sigma e_i (y_i - \hat{y}_i)$

$= \Sigma e_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$ as $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$

$= \Sigma e_i y_i - \hat{\beta}_1 \Sigma x_{1i} e_i - \hat{\beta}_2 \Sigma x_{2i} e_i$

$= \Sigma e_i y_i$

$= \Sigma (y_i - \hat{y}_i) y_i$

$= \Sigma y_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$

$= \Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

$\Sigma e_i^2 = \Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

Since $\Sigma x_{1i} e_i = 0$, $\Sigma x_{2i} e_i = 0$

where $\Sigma x_{1i}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$

$= \Sigma x_{1i} y_i - \hat{\beta}_1 \Sigma x_{1i}^2 - \hat{\beta}_2 \Sigma x_{1i} x_{2i}$

$= 0$

which follows from the first normal equation.

Similarly $\Sigma x_{2i} e_i = 0$

i.e., RSS = $\Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

since $\Sigma y_i^2 = $ TSS

ESS = $\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i$

$$R^2_{Y \cdot X_1 X_2} = \frac{ESS}{TSS} = \frac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i}{\Sigma y_i^2}$$

Example 3.3 [illegible]

(i) Find the value of R^2

(ii) Find the fitted regression equation

(b) Following Example 3.2,

(i) Find the value of R^2

[illegible]

Solution

[illegible]

[illegible]

We know that R^2 [illegible] $\frac{\hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i}{[illegible]}$

[illegible]

$R^2 = 0.$[illegible]

The estimated fitted regression equation is

[illegible]

[illegible]

[illegible] the estimated value of Example 3.2

We get [illegible]

Further we obtained $\hat{\beta}_1$ [illegible], $\hat{\beta}_2$ [illegible] 0.25 and $\hat{\beta}_3$ [illegible] 5.5

We thus get R^2 [illegible]

[illegible]

R^2 [illegible]

(ii) Thus the estimated regression equation is

$\hat{Y}_i = 23.75$ [illegible]

This equation suggests that years of experience with the firm is far more important than years of education (which actually has a negative sign). This equation says that we can predict that one more year of experience, after allowing for years of education (or holding it constant), results in an annual increase in salary of $550. This means that if we consider two persons with the same level of education, the one with one more year of experience can be expected to have a higher salary of [illegible]. Similarly, if we consider two persons with the same experience, the one with an education of one more year can be expected to have a lower annual salary of [illegible]

Here $R^2 = 0.995$ implies that out of [illegible] variation in salary of the employees [illegible] variation can be explained by the two explanatory variables X_2 and X_3 jointly.

4.7 R^2 and the Adjusted R^2

[illegible] will not decrease R^2.

[illegible]

Now TSS [illegible]

The RSS $\sum e_i^2$, however, depends on the [illegible] it is clear that as the number of X variables increases, $\sum e_i^2$ is likely to decrease (at least will not increase); hence R^2 will increase. [illegible] regression models with the same dependent variable [illegible] X variables, one should be very wary of choosing the model [illegible] R^2.

To compare two R^2 terms, one must take into account the number of X variables present in the model. This can be done readily if we consider [illegible] of determination,

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-(K+1))}{\sum y_i^2/(n-1)}$$

where K = number of explanatory variables and K + 1 = number of parameters in the model including the intercept term. In a three variable (with two explanatory variables) linear regression model $K = 2$ so $n - (K+1) = n - 3$.

The R^2 thus defined is known as adjusted R^2, denoted by $\bar{R}^2$. The term adjusted means adjusted for the degrees of freedom (d.f.) associated with the sums of squares of $\sum e_i^2$ and $\sum y_i^2$.

RSS $= \sum e_i^2$ has $n - (K+1)$ degrees of freedom in a model involving $(K+1)$ parameters, including the intercept term and TSS $= \sum y_i^2$ has $n - 1$ degrees of freedom.

Thus the adjusted R^2 can also be written as

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}_u^2}{S_Y^2}$$

It is thus clear that $\bar{R}^2$ and R^2 are related and we can express the relation as follows

where $\hat{\sigma}_u^2 = \sum e_i^2/(n-(K+1))$ is the residual variance and unbiased estimator of true σ_u^2 and

$$S_Y^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2 = \frac{1}{n-1}\sum y_i^2 = \text{sample variance of } Y$$

$$\sum y_i^2 = (n-1)S_Y^2 \text{ and } \sum y_i^2/(n-1) = S_Y^2$$

Since $\bar{R}$ [illegible]

[illegible]

[illegible]

[illegible] three-variable linear regression model, we have $\bar{R}^2$ [illegible] and R^2

[illegible]

[illegible] that for $k > 1$, $\bar{R}^2 < R^2$ which implies that as the number of X variables increases, the adjusted R^2 increases less than unadjusted R^2; $\bar{R}^2$ can be negative, although R^2 is necessarily non-negative. In case $\bar{R}^2$ turns out to be negative in an application, its value is taken as zero.

[illegible] should be noted that for a large $\bar{R}^2$ and R^2 [illegible] variables [illegible] the number of regressors [illegible] is large in relation to the sample [illegible] $\bar{R}^2$ will be much smaller than R^2 and can even assume negative values, in which case $\bar{R}^2$ should be interpreted as being equal to zero.

Note: Comparing Two R^2 values

It is crucial to note that in comparing two models on the basis of the coefficient of determination, whether adjusted or not, the sample size n and the dependent variable must be the same; the explanatory variables may take any form. Thus for the models

$$\log Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \qquad \text{.....} (A)$$

and $$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i \qquad \text{.....} (B)$$

the computed R^2 terms cannot be compared. The reason is that by definition R^2 measures the proportion of the variation in the dependent variable accounted for by the explanatory variables. Therefore in equation (A) R^2 measures the proportion of the variation in log Y explained by X_1 and X_2 whereas in equation (B) it measures the proportion of the variation in Y and hence the two are not the same thing. A change in log Y gives a relative or proportional change in Y whereas a change in Y gives an absolute change. Therefore var $\hat{Y}_i$/var Y_i is not equal to var(log [illegible])/var(log [illegible]). Thus the two coefficients of determination are not the same.

Example 3.4. (a) Following Example 3.1 find the value of Adjusted R^2.

(b) Following Example 3.2 find the value of Adjusted R^2.

Solution: (a) We know that in a three-variable linear regression model, adjusted R^2 is denoted by

$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-3}$. Here we see that following data of Example [illegible] R^2 [illegible]

$n = 5$

[illegible]

Thus the value of adjusted $R^2 = \bar{R}^2 = 0.13$

[illegible] that $\bar{R}^2 < R^2$ [illegible]

Following Example 3 [illegible] R^2 [illegible] $\frac{n-1}{n-3}$ [illegible] and $n = 5$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} = 1 - [illegible]$$

$$= 1 - 0.012 = 0.988$$

Adjusted $R^2 = \bar{R}^2 = 0.988$ which is smaller than unadjusted R^2 [illegible]

The value of adjusted $R^2 = \bar{R}^2$ can also be obtained by using the formula

$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-(K+1))}{\sum y_i^2/(n-1)}$$

From Example 3.2 we have obtained

$\sum e_i^2 = 1.5$, $\sum y_i^2 = 272$, $n = 5$, $K = 2$, $K + 1 = 3$

$$\bar{R}^2 = 1 - \frac{\frac{1.5}{5-3}}{\frac{272}{5-1}} = 1 - \frac{0.75}{68} = 1 - 0.0110 = 0.988$$

$$\bar{R}^2 = 0.988$$

3.8. Partial Correlation Coefficients and the Coefficient of Partial Determination

In the simple correlation analysis the coefficient of correlation r is used as a measure of the degree of linear association between two variables X and Y

$$[Y_i = \alpha + \beta X_i + u_i, \quad i = 1, 2, \ldots, n]$$

For three variable linear regression model $[Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \; i = 1, 2, \ldots, n]$ we can compute three correlation coefficients $r_{YX_1} = r_{12}$ (correlation coefficient between Y and X_1), $r_{YX_2} = r_{13}$ (correlation coefficient between Y and X_2) and $r_{X_1X_2} = r_{23}$ (correlation coefficient between X_1 and X_2)

These correlation coefficients are called gross or simple correlation coefficients or correlation coefficients of zero order and computed by the formula

$$r_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\Sigma(Y_i - \bar{Y})^2}} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}}$$

where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$

[illegible] them and with both of them.

[illegible]

Partial correlation coefficient

[illegible]

Partial correlation coefficient between [illegible] holding [illegible] constant

Partial correlation coefficient between [illegible] holding [illegible] constant

These partial correlations can be obtained from [illegible] simple [illegible]

[illegible]

The partial correlations defined above are called [illegible] By order we mean the number of secondary subscripts [illegible] correlation coefficient [illegible] and so on. [illegible]

In the two variable [illegible] meaning. It measures the degree of linear [illegible] dependent variable Y and the single explanatory variable [illegible] we need to pay careful attention to the [illegible] coefficient.

From the formula

[illegible] for example we [illegible] the following:

i) Even [illegible]

ii) If [illegible] and [illegible] will be negative whereas if they are of the opposite [illegible] will be positive.

iii) The terms [illegible] and [illegible] need not have the same sign.

iv) In the two variable case [illegible] property holds true for partial correlation coefficients [illegible]

[illegible] Similarly [illegible] and [illegible]

[illegible] and [illegible] [illegible] X_3 are uncorrelated

The [illegible] coefficient of partial determination [illegible] be interpreted as the proportion of the variation in Y not explained by the variable X_2 that has been explained by the inclusion of X_3 into the model. Conceptually it is similar to R^2 [illegible] coefficient of determination.

[illegible] [illegible] are found [illegible] R [illegible] and partial correlation coefficients

$$R = [illegible]$$

$$R^2 = [illegible]$$

$$R^2 = r_{12}^2 + (1 - r_{12}^2) r_{13.2}^2$$

It has been pointed out earlier that R^2 will not decrease if an additional explanatory variable is introduced into the model, which can be seen clearly from the equation $R^2 = r_{12}^2 + (1 - r_{12}^2) r_{13.2}^2$. This equation states that the proportion of the variation in Y explained by X_2 and X_3 jointly is the sum of two parts: the part explained by X_2 alone $= r_{12}^2$ and the part not explained by X_2 $(= 1 - r_{12}^2)$ times the proportion that is explained by X_3 after holding the influence of X_2 constant. Now $R^2 > r_{12}^2$ [illegible] as [illegible]. At worst, $r_{13.2}^2$ will be zero, in which case $R^2 = r_{12}^2$.

Example 3.5. Following Example 3.1

(i) Find the values of r_{12}, r_{13} and r_{23}.

(ii) Find the values of the partial regression coefficient $r_{12.3}$

(iii) Find the value of R^2 in terms of r_{12}, r_{13} and r_{23}.

Solution From the calculation table of Example 3.1 we have the following values $\Sigma x_{1i}^2 = 40$, $\Sigma x_{2i}^2 = 23.20$, $\Sigma y_i^2 = 24$, $\Sigma x_{1i} x_{2i} = 17$, $\Sigma x_{1i} y_i = 20$ and $\Sigma x_{2i} y_i = 3$

where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$ and $y_i = Y_i - \bar{Y}$

(i) Now by using the formula of r_{12}, r_{13} and r_{23} and putting the required values we can get the value of r_{12}, r_{13} and r_{23}

By definition, $r_{12} = r_{XY} = \dfrac{\text{cov}(X_1, Y)}{\sigma_X \sigma_Y} = \dfrac{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)^2}\sqrt{\frac{1}{n}\Sigma(Y_i - \bar{Y})^2}}$

$$= \frac{\frac{1}{n}\Sigma x_{1i} y_i}{\sqrt{\frac{1}{n}\Sigma x_{1i}^2}\sqrt{\frac{1}{n}\Sigma y_i^2}} = \frac{\Sigma x_{1i} y_i}{\sqrt{\Sigma x_{1i}^2}\sqrt{\Sigma y_i^2}} = \frac{20}{\sqrt{40 \times 24}} = 0.64$$

$$r_{12} = 0.64$$

Similarly, [illegible]

and [illegible]

$r_{12} = 0.64$, $r_{13} = -0.13$ and $r_{23} = 0.56$

(c) We know that [illegible]

$r_{12} = 0.56$

(d) We know that,

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{(0.64)^2 + (-0.13)^2 - 2 \times 0.64 \times (-0.13) \times 0.56}{1-(0.56)^2}$$

$$= \frac{0.4096 + 0.0169 + 0.093184}{0.6864} = \frac{0.5196}{0.6864} = 0.76$$

$R^2 = 0.76$

Example 3.6. Are the following data consistent? Give reasons.

(a) $r_{23} = 0.9$, $r_{13} = -0.2$, $r_{12} = 0.8$

(b) $r_{12} = 0.6$, $r_{23} = 0.9$, $r_{31} = -0.5$

Solution. From the above data set we will first calculate the value

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$ and will verify whether $0 \le R^2 \le 1$ or not.

(a) Here we see that

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{(0.8)^2 + (-0.2)^2 - 2 \times 0.8 \times (-0.2) \times 0.9}{1-(0.9)^2} = \frac{0.64 + 0.04 + 0.288}{0.19} = \frac{0.968}{0.19} = 5$$

$R^2 = 5$ which is not possible as $0 \le R^2 \le 1$.

Hence the given informations are not consistent.

(b) $$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$ as $r_{31} = r_{13} = -0.5$

$$= \frac{(0.6)^2 + (-0.5)^2 - 2 \times 0.6 \times (-0.5) \times 0.9}{1-(0.9)^2}$$

[illegible]

[illegible]

Example [illegible] standard errors and the [illegible]

$\sum(Y_i - \bar{Y})^2 = 66042.269$, $\sum(X_{1i} - \bar{X}_1)^2 = 84855.096$

$\sum(X_{2i} - \bar{X}_2)^2 = 280.000$, $\sum(Y_i - \bar{Y})(X_{1i} - \bar{X}_1) = 74778.346$

$\sum(Y_i - \bar{Y})(X_{2i} - \bar{X}_2) = 4250.900$, $\sum(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 4796.000$ and [illegible]

Solution The given results are related to a three variable linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, 2, \ldots, n$

where β_1 and β_2 are the partial regression coefficients. Assuming [illegible] x_{1i}, and $X_{2i} - \bar{X}_2 = x_{2i}$ we can obtain the OLS estimators [illegible] as follows

$$\hat\beta_1 = \frac{\sum x_{2i}^2 \sum x_{1i} y_i - \sum x_{1i} x_{2i} \sum x_{2i} y_i}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

$$= \frac{280.000 \times 74778.346 - 4796.00 \times 4250.900}{84855.096 \times 280.000 - (4796.00)^2}$$

$$= \frac{20937936.88 - 20387316.40}{23759426.88 - 23001616} = \frac{550620.48}{757810.88} = 0.7265$$

$$\hat\beta_1 = 0.7265$$

Similarly, $$\hat\beta_2 = \frac{\sum x_{1i}^2 \sum x_{2i} y_i - \sum x_{1i} x_{2i} \sum x_{1i} y_i}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

$$= \frac{84855.096 \times 4250.900 - 4796.00 \times 74778.346}{84855.096 \times 280.000 - (4796.00)^2}$$

$$= \frac{2073580.17}{757810.88} = 2.7362$$

$$\hat\beta_2 = 2.7362$$

Now we have to find out $SE(\hat\beta_1) = \sqrt{var(\hat\beta_1)}$ and $SE(\hat\beta_2) = \sqrt{var(\hat\beta_2)}$

We know that $$var(\hat\beta_1) = \frac{\sigma_u^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

and $$var(\hat\beta_2) = \frac{\sigma_u^2 \sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

[illegible]

[illegible]

Since [illegible]

[illegible]

Where [illegible]

Since [illegible]

$$ESS = \sum \hat{y}_i^2 = \hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}$$

[illegible]

Again $TSS = \sum y_i^2 =$ [illegible]

[illegible]

$\hat{\sigma}^2 =$ [illegible]

Now $var(\hat{\beta}_1) =$ [illegible]

$$SE(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = 0.0560$$

Similarly, $var(\hat{\beta}_2) =$ [illegible]

$$SE(\hat{\beta}_2) = \sqrt{var(\hat{\beta}_2)} = 0.8878$$

Now we have to find out the value of R^2 and adjusted $R^2 = \bar{R}^2$

We know that $R^2 = \dfrac{ESS}{TSS} = 1 - \dfrac{RSS}{TSS} = 1 - \dfrac{\sum e_i^2}{\sum y_i^2}$

[since $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \Rightarrow TSS = ESS + RSS$]

[illegible] Here n [illegible]

So R^2 and $\bar{R}^2$ are almost the same

4.9 Confidence Intervals and Hypothesis Testing in a Three Variable Multiple Linear Regression Model

We consider a multiple linear [illegible] given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, \ldots, n$$

Suppose $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}_u$ are the OLS [illegible] respectively.

We also know that

$\hat{\beta}_1 \sim N[\beta_1, \mathrm{var}(\hat{\beta}_1)]$ where $E(\hat{\beta}_1) = \beta_1$ and $\mathrm{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_{X_1}^2(1 - r_{X_1X_2}^2)}$. This is read as $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\mathrm{var}(\hat{\beta}_1)$

Similarly, $\hat{\beta}_2 \sim N[\beta_2, \mathrm{var}(\hat{\beta}_2)]$ where $E(\hat{\beta}_2) = \beta_2$

$$\text{and } \mathrm{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{n\sigma_{X_2}^2(1 - r_{X_1X_2}^2)}$$

$\hat{\beta}_0 \sim N[\beta_0, \mathrm{var}(\hat{\beta}_0)]$, where $E(\hat{\beta}_0) = \beta_0$

$$\text{and } \mathrm{var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} + \bar{X}_1^2\,\mathrm{var}(\hat{\beta}_1) + 2\bar{X}_1\bar{X}_2\,\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) + \bar{X}_2^2\,\mathrm{var}(\hat{\beta}_2)$$

and $u \sim N(0, \sigma_u^2)$ where $E(u) = 0$ and $\mathrm{var}(u) = \sigma_u^2$

$$\text{and } \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1 - r_{X_1X_2}^2)}$$

We are now interested in testing the following hypothesis

Case 1 We want to test the null hypothesis $H_0: \beta_0 = 0$ against the alternative hypothesis, either $H_1: \beta_0 \neq 0$ or $H_1: \beta_0 > 0$ or $H_1: \beta_0 < 0$

Since $\hat{\beta}_0 \sim N[\beta_0, \mathrm{var}(\hat{\beta}_0)]$

$$\text{Now } t \text{ or } Z = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim N(0, 1)$$

would be the appropriate test statistic where $SE(\hat{\beta}_0) = \sqrt{\mathrm{var}(\hat{\beta}_0)}$

When σ_u is unknown and it is replaced by its unbiased estimator [illegible].

$\left[\text{Since } E\left(\frac{\sum e_i^2}{n-3}\right) = \sigma_u^2\right]$, then the appropriate test statistic would be [illegible]

will follow a t distribution with d.f. = n − 3. Under H_0: β_0 = 0, the test statistic would be

$$t = \frac{\hat{\beta}_0}{[illegible]} = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$$

Nature of the Test

(i) For the alternative hypothesis $H_1: \beta_0 \neq 0$ the null hypothesis H_0 [illegible] be accepted at 100α% level of significance if for the given sample $-t_{\alpha/2,\, n-3} \le t \le t_{\alpha/2,\, n-3}$ and will be rejected otherwise, i.e., when [illegible].

(ii) For the alternative hypothesis $H_1: \beta_0 > 0$ the null hypothesis $H_0: \beta_0 = 0$ will be accepted if for the given sample $t \le t_{\alpha,\, n-3}$ and will be rejected otherwise, i.e., when $t > t_{\alpha,\, n-3}$.

(iii) For the alternative hypothesis $H_1: \beta_0 < 0$ the null hypothesis H_0 [illegible] be accepted if for the given sample $-t_{\alpha,\, n-3} \le t$ and will be rejected otherwise, i.e., when $t < -t_{\alpha,\, n-3}$. In each case α denotes the chosen level of significance. Usually α = 0.05 or 0.01.

Confidence interval for β_0:

As regards the problem of interval estimation of β_0 at 100α% level of significance the confidence limits to β_0 would be given by

$$\hat{\beta}_0 \pm t_{\alpha/2,\, n-3}\, SE(\hat{\beta}_0)$$

i.e. $P[-t_{\alpha/2,\, n-3} \le t \le t_{\alpha/2,\, n-3}] = 1 - \alpha$

or $P\left[-t_{\alpha/2,\, n-3} \le \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \le t_{\alpha/2,\, n-3}\right] = 1 - \alpha$

or $P\left[\hat{\beta}_0 - t_{\alpha/2,\, n-3}\, SE(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,\, n-3}\, SE(\hat{\beta}_0)\right] = 1 - \alpha$

Here (1 − α) is called the confidence coefficient.

Case 2. We want to test the null hypothesis $H_0: \beta_1 = 0$ against the alternative hypothesis $H_1: \beta_1 \neq 0$ or $H_1: \beta_1 > 0$ or $H_1: \beta_1 < 0$.

Since $\hat{\beta}_1 \sim N[\beta_1, \text{var}(\hat{\beta}_1)]$ where $E(\hat{\beta}_1) = \beta_1$

and $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_{X_1}^2 (1 - r_{X_1 X_2}^2)}$

Now t or Z = $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0, 1)$ would be the appropriate test statistic, where

[illegible]

[illegible] under $H_0: \beta_1 = 0$, the test statistic would be

$$[illegible]$$

Nature of the Test

(i) For the alternative hypothesis [illegible] $\neq 0$ the null hypothesis $H_0: \beta_1 = 0$ will be accepted if for the given sample [illegible] significance.

(ii) For the alternative hypothesis $H_1: \beta_1$ [illegible] be accepted if for the given sample [illegible]

(iii) For the alternative hypothesis $H_1: \beta_1$ [illegible] accepted if for the given sample [illegible] when [illegible]. In each case α (= 0.01 or 0.05) denotes the chosen level of significance.

Confidence interval for β_1

As regards the problem of interval estimation [illegible] the confidence limits to β_1 would be given by,

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-3} SE(\hat{\beta}_1)$$

i.e. $P[-t_{\alpha/2,\, n-3} \le t \le t_{\alpha/2,\, n-3}] = 1 - \alpha$

or, $P\left[-t_{\alpha/2,\, n-3} \le \dfrac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \le t_{\alpha/2,\, n-3}\right] = 1 - \alpha$

or, $P\left[\hat{\beta}_1 - t_{\alpha/2,\, n-3} SE(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,\, n-3} SE(\hat{\beta}_1)\right] = 1 - \alpha$

Here $(1 - \alpha)$ is the confidence coefficient.

Case 3 We want to test the null hypothesis $H_0: \beta_2 = 0$ against the alternative hypothesis, $H_1: \beta_2 \neq 0$ or $H_1: \beta_2 > 0$ or, $H_1: \beta_2 < 0$

Since $\hat{\beta}_2 \sim N[\beta_2, \text{var}(\hat{\beta}_2)]$ where $E(\hat{\beta}_2) = \beta_2$

$$\text{and } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{n\sigma_{X_2}^2(1 - r_{2,3}^2)}$$

Now t or $Z = \dfrac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \sim N(0,1)$ would be the appropriate test statistic where $SE(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)}$. When σ_u^2 is not known then it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-3)$ and the test statistic becomes $t = \dfrac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \sim t_{n-3}$

Under $H_0: \beta_2 = 0$ the test statistic would be

$$= \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t_{n-3}$$

Nature of the Test

(i) For the alternative hypothesis $H_1: \beta_2 \neq 0$ the null hypothesis H_0 will be accepted if for the given sample $|t| < t_{\alpha/2, n-3}$ and will be rejected otherwise.

(ii) For the alternative hypothesis $H_1: \beta_2 > 0$ the null hypothesis H_0 will be accepted if for the given sample $t < t_{\alpha, n-3}$ and will be rejected otherwise when $t > t_{\alpha, n-3}$.

(iii) For the alternative hypothesis $H_1: \beta_2 < 0$ the null hypothesis H_0 will be accepted if for the given sample $t > -t_{\alpha, n-3}$ and will be rejected otherwise [i.e. when $t < -t_{\alpha, n-3}$]

Confidence Interval for β_2:

As regards the problem of interval estimation of β_2 at $100\alpha\%$ level of significance the confidence limits to β_2 would be given by

$$\hat{\beta}_2 \pm t_{\alpha/2, n-3} SE(\hat{\beta}_2)$$

as $P[-t_{\alpha/2, n-3} < t < t_{\alpha/2, n-3}] = 1 - \alpha$

or $P\left[-t_{\alpha/2, n-3} < \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} < t_{\alpha/2, n-3}\right] = 1 - \alpha$

or $P\left[\hat{\beta}_2 - t_{\alpha/2, n-3} SE(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + t_{\alpha/2, n-3} SE(\hat{\beta}_2)\right] = 1 - \alpha$

where $(1 - \alpha)$ is the confidence coefficient.

Case 4 We want to test the null hypothesis $H_0: \beta_1 = \beta_2$ against the alternative hypothesis $H_1: \beta_1 \neq \beta_2$ or $H_1: \beta_1 > \beta_2$ or $H_1: \beta_1 < \beta_2$

Since $(\hat{\beta}_1 - \hat{\beta}_2) \sim N(\beta_1 - \beta_2, \text{var}(\hat{\beta}_1 - \hat{\beta}_2))$

where $E(\hat{\beta}_1 - \hat{\beta}_2) = E(\hat{\beta}_1) - E(\hat{\beta}_2) = \beta_1 - \beta_2$

and $\text{var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$

$$= \frac{\sigma_u^2}{n\sigma_{X_1}^2(1 - r_{X_1X_2}^2)} + \frac{\sigma_u^2}{n\sigma_{X_2}^2(1 - r_{X_1X_2}^2)} + \frac{2\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1 - r_{X_1X_2}^2)}$$

as $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1 - r_{X_1X_2}^2)} = \frac{\sigma_u^2}{n(1 - r_{X_1X_2}^2)}\left[\frac{1}{\sigma_{X_1}^2} + \frac{1}{\sigma_{X_2}^2} + \frac{2r_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}\right]$

The appropriate test statistic would be given by

$$t \text{ or } Z = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim N(0, 1)$$

where $SE(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_1 - \hat{\beta}_2)}$

[illegible] is unknown and [illegible]

[illegible]

[illegible]

[illegible] would be [illegible]

Nature of the Test

(i) For the alternative hypothesis H_1 [illegible] $\beta_1 = \beta_2$ will be accepted if for the given sample [illegible] and will be rejected otherwise.

(ii) For the alternative hypothesis H_1 [illegible] be accepted if for the given sample [illegible] when t [illegible] $t_{\alpha, n-3}$)

(iii) For the alternative hypothesis H_1 [illegible] be accepted if for the given sample [illegible] i.e. when t [illegible] $t_{\alpha, n-3}$). In each case α denotes the [illegible]

Confidence interval of $(\beta_1 - \beta_2)$:

At 100 α% level of significance, the confidence interval [illegible] β_2) would be given by

$$(\hat{\beta}_1 - \hat{\beta}_2) \pm t_{\alpha/2, n-3} SE(\hat{\beta}_1 - \hat{\beta}_2)$$

i.e., $P(-t_{\alpha/2, n-3} \leq t \leq t_{\alpha/2, n-3}) = 1 - \alpha$

or, $$P\left[-t_{\alpha/2, n-3} \leq \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \leq t_{\alpha/2, n-3}\right] = 1 - \alpha$$

or, $$P\left[(\hat{\beta}_1 - \hat{\beta}_2) - t_{\alpha/2, n-3} SE(\hat{\beta}_1 - \hat{\beta}_2) \leq (\beta_1 - \beta_2) \leq (\hat{\beta}_1 - \hat{\beta}_2) + t_{\alpha/2, n-3} SE(\hat{\beta}_1 - \hat{\beta}_2)\right]$$

$= 1 - \alpha$

where $(1 - \alpha)$ is the confidence coefficient.

Case 5 : Confidence interval for σ_u^2:

Under the normality assumption, the variable

$$\chi^2 = \frac{RSS}{\sigma_u^2} = \frac{\Sigma e_i^2}{\sigma_u^2} = (n-3)\frac{\hat{\sigma}_u^2}{\sigma_u^2}$$

follows a χ^2 (chi-square) distribution with d.f. = $n - 3$, [where $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-3)$, is an unbiased estimator of σ_u^2]

Therefore we can use χ^2 to establish a confidence interval for σ_u^2.

At the α level of significance the confidence interval for σ_u^2 can be given by

[illegible] and [illegible] where χ^2 values are taken from the table with d.f. = (n − 3)

[illegible]

[illegible]

$$P\left[(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2}} < \sigma_u^2 < (n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

where (1 − α) is the confidence coefficient.

Example 3.8. The following table contains observations on the quantity demanded (Y) of a certain commodity, its price (X_1) in $ and consumer's income (X_2) in $.

Y	100	75	80	70	50	65	90	100	110	60
X_1	5	7	6	6	8	7	5	4	3	9
X_2	1000	600	1200	500	300	400	1300	1100	1300	300

Assume a linear regression equation of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, 3, \ldots, 10$$

(i) Find the OLS estimators of β_0, β_1 and β_2 i.e. $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$

(ii) Find R^2 and adjusted R^2 ($\bar{R}^2$)

(iii) Find $\operatorname{var}(\hat{\beta}_0)$, $\operatorname{var}(\hat{\beta}_1)$ and $\operatorname{var}(\hat{\beta}_2)$

(iv) Find $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_2)$

(v) Write the regression results in the summary form.

(vi) Test $H_0: \beta_0 = 0$ against $H_1: \beta_0 \neq 0$ and find 95% and 99% confidence intervals for β_0.

(vii) Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ and find 95% and 99% confidence intervals for β_1.

(viii) Test $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$ and find 95% and 99% confidence intervals for β_2.

(ix) Test $H_0: \beta_1 = \beta_2$ against $H_1: \beta_1 \neq \beta_2$ and find 95% and 99% confidence intervals for $(\beta_1 - \beta_2)$.

(x) Construct 95% and 99% confidence intervals of σ_u^2.

Calculation Table 3.3

n	Y	X_1	X_2	$y_i = Y_i - \bar{Y}$	$x_{1i} = X_{1i} - \bar{X}_1$	$x_{2i} = X_{2i} - \bar{X}_2$	y_i^2	x_{1i}^2	x_{2i}^2	$x_{1i}y_i$	$x_{2i}y_i$	$x_{1i}x_{2i}$
1	100	5	1000	20	1	200	400	1	40,000	20	4000	200
2	75	7	600	5	1	200	25	1	40,000	5	1000	-200
3	80	6	1200	0	0	400	0	0	160,000	0	0	0
4	70	6	500	-10	0	300	100	0	90,000	0	3000	0
5	50	8	300	30	2	500	900	4	250,000	-60	15000	1000
6	65	7	400	-15	1	-400	225	1	160,000	15	6000	400
7	90	5	1300	10	-1	500	100	[illegible]	250,000	[illegible]	5000	500
8	100	4	1100	20	2	300	400	4	90,000	40	6000	600
9	110	3	1300	30	3	500	900	9	250,000	[illegible]	[illegible]	[illegible]
10	60	9	300	20	3	500	400	9	250,000	[illegible]	[illegible]	[illegible]
$n = 10$	$\sum Y_i = 800$	$\sum X_{1i} = 60$	$\sum X_{2i} = 8000$	$\sum y_i = 0$	$\sum x_{1i} = 0$	$\sum x_{2i} = 0$	$\sum y_i^2 = 3450$	$\sum x_{1i}^2 = 30$	$\sum x_{2i}^2 = 1{,}580{,}000$	[illegible]	$\sum x_{2i}y_i = 65000$	$\sum x_{1i}x_{2i} = 5500$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{800}{10} = 80, \quad \bar{X}_1 = \frac{\sum X_{1i}}{n} = \frac{60}{10} = 6 \text{ and } \bar{X}_2 = \frac{\sum X_{2i}}{n} = \frac{8000}{10} = 800$$

Solution

The [illegible] estimators are [illegible]

[illegible]

Similarly $\hat{\beta}_2$ [illegible]

[illegible]

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ can be obtained from the equation

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$

or, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ where [illegible]

= [illegible]

= [illegible] $\hat{\beta}_0$ = [illegible]

Thus we have $\hat{\beta}_0$ = [illegible], $\hat{\beta}_1$ = [illegible] and $\hat{\beta}_2$ = [illegible]

ii. We know that $R^2 = \frac{ESS}{TSS}$ = [illegible]

= [illegible] = 0.894

R^2 = 0.894. This means that price and income can jointly explain 89.4% variation in demand out of total variation of 100%.

Now, adjusted $R^2 = \bar{R}^2$ = [illegible]. Here n = [illegible]

$\bar{R}^2$ = [illegible] (1 − 0.894) [illegible] = 0.862 [illegible]

R^2 = 0.894 and adjusted R^2 = $\bar{R}^2$ = 0.8627

iii. We know that $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 \sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$

Here σ_u^2 is unknown and it is replaced by its unbiased estimator $\hat{\sigma}_u^2$ [illegible]

We know that $R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y^2} = 1 - \frac{RSS}{TSS}$

[illegible]

[illegible]

Now var [illegible]

$= \frac{[illegible]}{30 \times 580,000 - (-5900)^2} = \frac{[illegible]}{[illegible]}$

var$(\hat{\beta}_1)$ = 6.55

Again var$(\hat{\beta}_2) = \frac{\hat{\sigma}_u^2 \, [illegible]}{\Sigma x_{1i}^2 \, \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$= \frac{52.24 \times 30}{30 \times 580,000 - (-5900)^2} = \frac{[illegible]}{[illegible]} = 0.000124$

$\text{var}(\hat{\beta}_2) = 0.000124$

Again, $\text{var}(\hat{\beta}_0) = \frac{\hat{\sigma}_u^2}{n} + \bar{X}_1^2 \, \text{var}(\hat{\beta}_1) - 2\bar{X}_1 \bar{X}_2 \, \text{cov}(\hat{\beta}_1, \hat{\beta}_2) + [illegible] \, \text{var}(\hat{\beta}_2)$

We know that $\hat{\sigma}_u^2 = 52.24$, $n = 10$, $\bar{X}_1 = 6$, $\bar{X}_2 = 800$, var$(\hat{\beta}_1)$ = 6.55, var$(\hat{\beta}_2)$ = 0.000124

and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\hat{\sigma}_u^2 \Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2 \, \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$= \frac{52.24 \times (-5900)}{30 \times 580,000 - (-5900)^2} = \frac{308216}{12590000} = 0.0245$

$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.0245$. We now put these values in the expression of $\hat{\beta}_0$ and we get,

$\text{var}(\hat{\beta}_0) = \frac{52.24}{10} + (6)^2 \times 6.55 + 2 \times 6 \times 800 \times 0.0245 + (800)^2 \times [illegible]$

$= 5.224 + 235.8 + 235.2 + 79.36 = 555.58$

(iv) We know that $SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$

$SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{6.55} = 2.5592$

[illegible]

[illegible]

The regression results in summary form

[illegible]

[illegible]

$R^2 = 0.194$, Adjusted $R^2 = \bar{R}^2 = 0.5617$

We want to test the null hypothesis $H_0: \beta_0 = 0$ against the alternative [illegible]. The appropriate test statistic under $H_0: \beta_0 = 0$ would be

$$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \sim t_n$$

Here t (observed) $= \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} = \frac{111.70}{[illegible]} = 4.719$

The null hypothesis $H_0: \beta_0 = 0$ will be accepted if [illegible] for the given sample [illegible] and will be rejected otherwise.

When α [illegible] $t_{0.025, 10}$ [illegible] $t_{0.025,7} = 2.365$

$t_{0.025,7} = 2.365$ (Table value)

Thus we see that t (observed) $= 4.719$ does not lie in the interval -2.365 and 2.365 [illegible]. Hence $H_0: \beta_0 = 0$ is rejected and $H_1: \beta_0 \neq 0$ is accepted at 5% level of significance.

Similarly, when $\alpha = 0.01$ [illegible] $= 3.499$. Here we see that t (observed) $= 4.719$ does not lie in the interval -3.499 and 3.499. Hence $H_0: \beta_0 = 0$ is rejected and $H_1: \beta_0 \neq 0$ is accepted at 1% level of significance.

We know that $100(1 - \alpha)\%$ confidence interval of β_0 would be

$$P\left[-t_{\alpha/2,n} \le t \le t_{\alpha/2,n}\right] = 1 - \alpha$$

$$\text{or, } P\left[-t_{\alpha/2,n} \le \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \le t_{\alpha/2,n}\right] = 1 - \alpha$$

$$\text{or, } P\left[\hat{\beta}_0 - t_{\alpha/2,n}\, SE(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n}\, SE(\hat{\beta}_0)\right] = 1 - \alpha$$

when $\alpha = 0.05$, $P[\hat{\beta}_0 - t_{0.025,7} \times 23.570 \le \beta_0 \le \hat{\beta}_0 + t_{0.025,7} \times 23.570] = 1 - 0.05 = 0.95$

or, $P[111.70 - 2.365 \times 23.570 \le \beta_0 \le 111.70 + 2.365 \times 23.570] = 0.95$

or, $P[55.957 \le \beta_0 \le 166.743] = 0.95$

95% confidence intervals of β_0 are 55.957 and 166.743

Similarly, when $\alpha = 0.01$, then $100(1 - \alpha)\% = 99\%$

So 99% confidence intervals of β_0 would be

$$\hat{\beta}_0 \pm t_{\alpha/2,n-3} SE(\hat{\beta}_0)$$

i.e., $\hat{\beta}_0 \pm t_{0.005,7} SE(\hat{\beta}_0)$

or, $111.70 \pm 3.499 \times 23.570$

or, 111.70 ± 82.4714

or, 29.2286 and 194.1714

So 99% confidence intervals of β_0 are 29.2286 and 194.1714

(vi) To test the null hypothesis $H_0: \beta_1 = 0$ against the alternative $H_1: \beta_1 \neq 0$ the test statistic under $H_0: \beta_1 = 0$ would be $t(\text{observed}) = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-3}$

Here $t(\text{observed}) = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-7.19}{2.5592} = -2.8094$

Now, $H_0: \beta_1 = 0$ will be accepted if for the given sample $-t_{\alpha/2,n-3} \le t(\text{observed}) \le t_{\alpha/2,n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,n-3} = t_{0.025,(10-3)} = t_{0.025,7} = 2.365$

and when $\alpha = 0.01$, $t_{\alpha/2,n-3} = t_{0.025,7} = 3.499$

Here we see that $t(\text{observed}) = -2.8094$ does not lie in the interval -2.365 and 2.365 and hence $H_0: \beta_1 = 0$ is rejected at 5% level of significance. But $t(\text{observed}) = -2.8094$ lies in the interval -3.499 and 3.499 and hence $H_0: \beta_1 = 0$ is accepted at 1% level of significance.

(vii) $100(1-\alpha)$% confidence limits to β_1 would be

$$\hat{\beta}_1 \pm t_{\alpha/2,n-3} SE(\hat{\beta}_1)$$

when $\alpha = 0.05$, then 95% confidence limits to β_1 would be

$\hat{\beta}_1 \pm t_{0.025,7} SE(\hat{\beta}_1)$

or, $-7.19 \pm 2.365 \times 2.5592$

or, -7.19 ± 6.0525 or, -13.245 and -1.135

So, 95% confidence limits to β_1 are -13.245 and -1.135

Similarly, when $\alpha = 0.01$ then 99% confidence limits to β_1 would be

$\hat{\beta}_1 \pm t_{0.025,7} SE(\hat{\beta}_1)$

or, $-7.19 \pm 3.499 \times 2.5592$

or, -7.19 ± 8.9546 or -16.1446 and 1.7646

So, 99% confidence limits to β_1 are -16.1446 and 1.7646

(viii) To test the null hypothesis $H_0: \beta_2 = 0$ against the alternative $H_1: \beta_2 \neq 0$ the appropriate test statistic under $H_0: \beta_2 = 0$ would be

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t_{n-3}$$

[illegible]

[illegible]

and will be rejected otherwise.

When [illegible]

and when [illegible]

Here we see that [illegible] observed [illegible]

which lies within both the critical values [illegible] 2.365 and [illegible] 3.499 [illegible] is not significant both at 5% and 1% levels of significance and hence [illegible] confidence intervals of β_2 would be

$$\hat{\beta}_2 \pm t_{\alpha/2,\,7}\ SE(\hat{\beta}_2)$$

when α = [illegible] then 95% confidence intervals of β_2 would be

$$\hat{\beta}_2 \pm [illegible]\ SE(\hat{\beta}_2)$$

or $0.0143 \pm 2.365 \times 0.011$

or [illegible] 0.0262 or [illegible] and [illegible]

So 95% confidence intervals of β_2 are [illegible] and [illegible]

When α = [illegible] then [illegible] 99% confidence intervals of β_2 would be

$$\hat{\beta}_2 \pm t_{0.005,\,7}\ SE(\hat{\beta}_2)$$

or [illegible] ± [illegible] 3.499 × 0.011

or [illegible] 0.0385 or [illegible] 0.0245 and 0.053[illegible]

So 99% confidence intervals of β_2 are [illegible] 0.0245 and 0.053[illegible]

(ii) To test the null hypothesis $H_0: \beta_1 = \beta_2$ against the alternative $H_1: \beta_1 \neq \beta_2$ the appropriate test statistic under $H_0: \beta_1 = \beta_2$ would be

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-2}$$

$$\text{Now } t\ (\text{observed}) = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

Since $\hat{\beta}_1 =$ [illegible], $\hat{\beta}_2 = 0.0143$, $\text{var}(\hat{\beta}_1) = 6.55$, $\text{var}(\hat{\beta}_2) = 0.000124$ and

$$SE(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2)}$$

$$= \sqrt{6.55 + 0.000124 - 2 \times 0.0245}$$

$$= \sqrt{6.55 + 0.000124 - 0.049} = \sqrt{6.501} = 2.549$$

[illegible] observed [illegible]

Now [illegible] will be accepted if [illegible] will be rejected otherwise [illegible]

When $\alpha = 0.05$, $t_{\alpha/2, n-3} = t_{0.025, 7} = 2.365$

and when [illegible]

Here we see that t observed [illegible] does not [illegible] and hence $H_0: \beta_1 = \beta_2$ is rejected and $H_1: \beta_1 \neq \beta_2$ is accepted at 5% [illegible] significance.

But we see that t observed [illegible] 2.8263 lies in the interval [illegible] the null hypothesis $H_0: \beta_1 = \beta_2$ is accepted at [illegible] of significance.

Now $100(1-\alpha)\%$ confidence intervals of $(\beta_1 - \beta_2)$ would be

$$(\hat{\beta}_1 - \hat{\beta}_2) \pm t_{\alpha/2, n-3}\, SE(\hat{\beta}_1 - \hat{\beta}_2)$$

When $\alpha = 0.05$ then $100(1-\alpha)\% = 95\%$ confidence intervals of $(\beta_1 - \beta_2)$ would be

$$(\hat{\beta}_1 - \hat{\beta}_2) \pm t_{0.025, 7}\, SE(\hat{\beta}_1 - \hat{\beta}_2)$$

or, $(-7.[illegible]9 - 0.0[illegible]43) \pm 2.365 \times 2.549$

or, -7.2043 ± 6.0283 or, -13.2326 and -1.176

So 95% confidence intervals of $(\beta_1 - \beta_2)$ are -13.2326 and -1.176

When $\alpha = 0.01$ then $100(1-\alpha)\% = 99\%$ confidence intervals of $(\beta_1 - \beta_2)$ would be

$$(\hat{\beta}_1 - \hat{\beta}_2) \pm t_{\alpha/2, n-3}\, SE(\hat{\beta}_1 - \hat{\beta}_2)$$

or $(-7.[illegible]9 - 0.0[illegible]3) \pm t_{0.005, 7} \times 2.549$

or $-7.2043 \pm 3.499 \times 2.549$

or, -7.2043 ± 8.9189 or, -16.1232 and 1.7146

So, 99% confidence intervals of $(\beta_1 - \beta_2)$ are -16.1232 and 1.7146

(i) We have to construct 95% and 99% confidence intervals of σ_u^2

We know that $100(1-\alpha)\%$ confidence intervals of σ_u^2 would be given by

$$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2, n-3}} \text{ and } (n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2, n-3}}$$ where χ^2 values are taken from the table

with d.f = $(n-3)$

When $\alpha = 0.05$ $\chi^2_{\alpha/2, n-3} = \chi^2_{0.025, 7} = 16.013$

and $\chi^2_{1-\alpha/2, n-3} = \chi^2_{0.975, 7} = 1.690$

[illegible]

$$[illegible]$$

and $$[illegible] = \frac{52.24}{[illegible]} = \frac{7 \times 52.24}{[illegible]} = 216.178$$

[illegible]

and [illegible]

So 95% confidence interval of [illegible] would be

$$[illegible]$$

and $$[illegible]$$

3.10 Analysis of Variance (ANOVA) in a Multiple Linear (Three Variable) Regression Model

Yet another item that is often presented in connection with the three variable linear regression model is the **analysis of variance**. This is the break down of the total sum of squares (TSS) into explained sum of squares (ESS) and the residual sum of squares (RSS).

The estimated three variable linear regression line, where the regression equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ [illegible] is given by $\hat{Y} = \hat\beta_0 + \hat\beta_1 X_{1i}$ [illegible] and $\bar{Y} = \hat\beta_0 + \hat\beta_1 \bar{X}_1 + \hat\beta_2 \bar{X}_2$

Taking deviations from mean we have

$$\hat{Y}_i - \bar{Y} = \hat\beta_0 + \hat\beta_1 X_{1i} + \hat\beta_2 X_{2i} - \hat\beta_0 - \hat\beta_1 \bar{X}_1 - \hat\beta_2 \bar{X}_2$$

or $$\hat{y}_i = \hat\beta_1 (X_{1i} - \bar{X}_1) + \hat\beta_2 (X_{2i} - \bar{X}_2)$$

or $$\hat{y}_i = \hat\beta_1 x_{1i} + \hat\beta_2 x_{2i} \quad \text{where } x_{1i} = X_{1i} - \bar{X}_1 \text{ and } x_{2i} = X_{2i} - \bar{X}_2$$

Now, error of estimate $e_i = y_i - \hat{y}_i = y_i - \hat\beta_1 x_{1i} - \hat\beta_2 x_{2i}$ [illegible]

Now, $\Sigma y_i^2 = \Sigma(\hat{y}_i + e_i)^2 = \Sigma \hat{y}_i^2 + 2\Sigma \hat{y}_i e_i + \Sigma e_i^2$

$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$ as $\Sigma \hat{y}_i e_i = 0$ by assumption

i.e. TSS = ESS + RSS

with d.f. $(K-1) = (K$ [illegible]

$K = 3$ as there are three [illegible]

regression model [illegible]

[illegible]

Thus we are [illegible] the unexplained [illegible] variation [illegible] analysis [illegible] variables [illegible] type of table of the regression [illegible] significance of the regression equation [illegible]

ANOVA TABLE

Table 3.4

Source of variation	Sum of squares	Degrees of freedom	Mean square	F (computed)	F (tabulated)
Explained (between)	$ESS = \hat\beta_1 \Sigma x_{1i} y_i + \hat\beta_2 \Sigma x_{2i} y_i$	$K - 1$	$\frac{ESS}{K-1} = MSE$	$F^* = \frac{MSE}{MSR}$ with d.f. $(K-1, n-K)$	
Residual (within)	$RSS = \Sigma e_i^2$	$n - K$	$\frac{RSS}{n-K} = MSR$		
Total	$TSS = \Sigma y_i^2$	$n - 1$			

In case of a three variable linear regression model there are three parameters and hence $K = 3$.

The test aims at finding out whether the explanatory variables (X_1 and X_2) do actually have any significant influence on the dependent variable (Y). Formally the test of the overall significance of the regression implies testing the null hypothesis $H_0: \beta_1 = \beta_2 = 0$ against the alternative hypothesis H_1: not all β_i are zero. We may use the test statistic

$$F^* = \frac{MSE}{MSR} = \frac{\Sigma \hat{y}^2/(K-1)}{\Sigma e_i^2/(n-K)}$$

$$= \frac{\Sigma \hat{y}_i^2/(3-1)}{\Sigma e_i^2/(n-3)} = \frac{\Sigma \hat{y}^2/2}{\Sigma e_i^2/(n-3)}$$ with d.f. 2, $n - 3$ (Here $K = 3$)

Now we have to compare F^*_{cal} with the table value of F with d.f. $(2, n-3)$. If it is found that $F^* > F_{\alpha, 2, n-3}$ (Table value) we reject the null hypothesis at $100\alpha\%$ level of significance ($\alpha = 0.01$ or 0.05 usually), i.e. we accept that the regression is significant and not all β's are zero.

If $F^* < F_{0.05}$, we accept the null hypothesis that [illegible] β_2 [illegible] the regression is not significant.

Note: Relation between R^2 and F

There is an intimate relationship between the coefficient [illegible] multiple determination R^2 and the F test used in the analysis of variance. Assuming the normal [illegible] for the disturbances u_i and the null hypothesis that [illegible] β_3 [illegible] we see that

$$F^* = \frac{ESS/[illegible]}{RSS/(n-K)} = \frac{ESS/[illegible]}{RSS/(n-3)}$$

is distributed as the F distribution with [illegible] and n [illegible] K [illegible] here [illegible] three parameters including the constant intercept term [illegible] three-variable multiple regression model.

Now we can write $F = F^* = \dfrac{ESS/(K-1)}{RSS/(n-K)}$

$$= \frac{n-K}{K-1}\cdot\frac{ESS}{RSS} = \frac{n-K}{K-1}\cdot\frac{ESS}{TSS-ESS}$$

$$= \frac{n-K}{K-1}\cdot\frac{ESS/TSS}{1-(ESS/TSS)} = \frac{n-K}{K-1}\cdot\frac{R^2}{1-R^2} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

$$F^* = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

It should be noted that here K = number of parameters in the multiple regression model [illegible] K = [illegible] when there are two explanatory variables. When $R^2 = 0$, F^* [illegible]. The larger the R^2, the greater the F^* value.

In the limit, when $R^2 = 1$, F is infinite. Thus the F test, which is a measure of the overall significance of the estimated regression, is also a test of significance of R^2. In other words, testing the null hypothesis $H_0: \beta_2 = \beta_3 = 0$ is equivalent to test the null hypothesis that population R^2 is zero. The ANOVA table can also be written expressed in terms of R^2 as shown below.

ANOVA TABLE in terms of R^2

Table 3.5

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean sum of squares (MS)	F	
				Observed	Tabulated
Explained (between)	$ESS = \Sigma\hat{y}^2$ $= R^2\,\Sigma y_i^2$	$K-1$	$ESS/(K-1)$ $= MSE$	$F = \dfrac{MSE}{MSR}$ with d.f. $= (K-1),\ (n-K)$	
Residual (within)	$RSS = \Sigma e_i^2 =$ $= (1-R^2)\,\Sigma y_i^2$	$n-K$	$RSS/(n-K)$ MSR		
Total	$TSS = \Sigma y_i^2$	$n-1$			

[illegible] variable linear regression [illegible]

Example 3.9 [illegible] H_1: β_1 and β_2 are not zero.

Solution [illegible]

[illegible] $\frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$ and $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$ [illegible]

[illegible] ANOVA Table

ANOVA TABLE relating to demand for a commodity

Table 3.8

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean sum of squares (MSS)	[illegible]	[illegible]
Explained (between)	$ESS = \Sigma\hat{y}_i^2$ 3086.5	$K - 1 = 2$	$MSE = \frac{ESS}{K-1}$ 1543.25	$F = F^* = \frac{MSE}{MSR}$	[illegible]
Residual (within)	$RSS = \Sigma e_i^2 =$ 363.5	$n - K = 7$	$MSR = \frac{RSS}{n-K}$ 51.92	with d.f. $(K - 1, n - K)$ 29.72	[illegible]
Total	$\Sigma y^2 = 3450$	$n - 1 = 9$			

Here the sample size, $n = 10$, number of parameters $K = 3$

$K - 1 = 2$ and $n - K = 10 - 3 = 7$

From Example 3.8 we have obtained the results

$$ESS = \Sigma\hat{y}_i^2 = \hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i = 3086.5$$

$$RSS = \Sigma e_i^2 = 363.5 \text{ and } TSS = \Sigma y_i^2 = 3450$$

Now the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ will be rejected if for the given sample

$$F = F^* \text{(observed)} = \frac{MSE}{MSR} \text{ [with d.f } (K - 1) = 2 \text{ and } (n - K) = 7]$$

is greater than the table value of F with d.f. $(K - 1) = 2$ and $(n - K) = 7$. From the table value we see that $F_{0.05,\,2,7} = 7.74$ and $F_{0.01,\,2,7} = 9.55$.

Here we see that F (observed) $= F^* = 29.72$ and $F_{0.05,\,2,7} = 7.74$

$F^* > F_{0.5\,2,7}$

So, at 5% level of significance the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ will be rejected for the given sample.

We also see that $F^* = 29.72 > F_{0.01,\,2,7} = 9.55$. This means that at 1% level of significance the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ will be rejected for the given sample.

Thus both at 1% and 5% levels of significance we may claim that [illegible] coefficients of the regression equation are not zero.

[illegible] should be noted that we can also construct the ANOVA [illegible] in [illegible] Following Example 3.8 we are showing the ANOVA table below in terms [illegible] R^2.

ANOVA TABLE in terms of R^2

Table 3.7

Sources of variation	Sum of squares SS	Degrees of freedom (d.f.)	Mean sum of squares (MS)	F	
				Observed	Tabulated
Explained (between)	$ESS = R^2\Sigma y^2$ $= 0.894 \times 3450$ $= 3084.30$	$K - 1 = 2$	$ESS/(K-1)$ $= MSE$ $= \frac{3084.30}{2}$ $= 1542.15$	$F = F^*$ $= \frac{MSE}{MSR}$ $= \frac{1542.15}{52.242}$ $= 29.52$	$F_{0.05}$ $= 4.74$
Residual (within)	RSS $= (1 - R^2)\Sigma y^2$ $= (1 - 0.894) \times 3450$ $= 365.70$	$n - K = 7$	$RSS/(n-K)$ $= MSR$ $= \frac{365.70}{7}$ $= 52.242$		$F_{0.01}$ $= 9.55$
Total	TSS $= \Sigma y^2 = 3450$	$n - 1 = 9$	–		

From Example 3.8 we have seen that $n = 10$, $K = 3$, $\Sigma y^2 = 3450$ and $R^2 = 0.894$. Here also we see that $F = F^* = 29.52 > F_{0.05, 2, 7} = 4.74$ and $F^* = 29.52 > F_{0.01, 2, 7} = 9.55$.

Thus the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is rejected both at 1% and 5% levels of significance.

3.11. The Cobb-Douglas Production Function : More on Functional Form

In Section 2.18 we showed how with appropriate transformations we can convert non-linear relationships into linear ones so that we can work within the framework of classical linear regression model. We consider the Cobb-Douglas Production function which shows a three variable non-linear relation. The Cobb-Douglas Production function, in its stochastic form, may be expressed as

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} U_i$$

where Y = output, X_1 = labour input, X_2 = capital input, U = Stochastic disturbance term, β_0 = constant technological parameter.

Taking Log on both sides of the [illegible]

$$\text{Log } Y = \log \beta_0 \ [illegible]$$

or $y = \alpha + \beta_1 x_{1i}$ [illegible]

This is a [illegible]

[illegible] and [illegible]

We can now apply the [illegible] parameters. The [illegible] using the OLS estimators of the Cobb-Douglas production function are

(i) β_1 is the partial [illegible] with respect of labour input i.e.

$$\frac{\partial \log Y}{\partial \log X_1} = \beta_1$$

(ii) Likewise β_2 is the partial [illegible] input

$$\text{i.e. } \frac{\partial \log Y}{\partial \log X_2} = \beta_2$$

(iii) The sum $(\beta_1 + \beta_2)$ gives information about the [illegible] displays IRS, CRS and DRS according as $\beta_1 + \beta_2 \gtrless 1$.

Example 3.10. A production function is specified as $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} U_i$ where $i = 1, 2, \ldots, n$

Y = output, X_1 = labour input, X_2 = capital input, U = Stochastic disturbance term, n = sample size. The corresponding Log-Linear form of the production function is given as

$$\text{Log } Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log U_i$$

$$\text{or } \quad y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \quad u_i \sim N(0, \sigma_u^2)$$

On the basis of a sample size of 23 the following results are given: $\hat{\alpha} = 4.0$, $\hat{\beta}_1 = 0.7$, $\hat{\beta}_2 = 0.2$, $RSS = 1.4$, $TSS = 10$, $\text{var}(\hat{\alpha}) = 0.6084$, $\text{var}(\hat{\beta}_1) = \text{var}(\hat{\beta}_2) =$ [illegible] 1.35

(i) Write the estimated regression equation.

(ii) Find the value of R^2

(iii) Find $SE(\hat{\alpha})$, $SE(\hat{\beta}_1)$, and $SE(\hat{\beta}_2)$

(iv) Find $\hat{\sigma}_u^2$

(v) Find the 95% confidence intervals for α, β_1, β_2 and σ_u^2

(vi) Test the hypothesis $\beta_1 = 1.0$ and $\beta_2 = 0$ separately at the 5% significance level

Solution : (i) The estimated regression equation can be written as

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

$$\text{or, } \quad \hat{y}_i = 4.0 + 0.7 x_{1i} + 0.2 x_{2i}$$

[illegible] The value of multiple coefficient of determination [illegible] given by R^2 [illegible]

where $ESS = TSS - RSS$ [illegible]

We know that $SE(\hat{\alpha}) = \sqrt{\text{variance}}$ [illegible]

Similarly, $SE(\hat{\beta}_1) = \sqrt{\text{variance}}$ [illegible]

and $SE(\hat{\beta}_2) = \sqrt{\text{variance}}$ [illegible]

(…) We have to find out the value of OLS estimator of σ_u i.e. $\hat{\sigma}_u$

We know that $\hat{\sigma}_u^2 = \frac{\sum e^2}{n-3} = \frac{RSS}{n-3}$

Here $RSS = 1.4$ and $n = 23$

$$\hat{\sigma}_u^2 = \frac{RSS}{n-3} = \frac{1.4}{23-3} = \frac{1.4}{20} = 0.07$$

(…) We have to find out the 95% confidence intervals for α, β_1, β_2 and σ_u. Using the t distribution with d.f. $= (n-3) = (23-3) = 20$, we can get the 95% confidence intervals for α, β_1, and β_2 as

For α: $\hat{\alpha} \pm t_{0.025,20} SE(\hat{\alpha})$ [illegible] 2.086 [illegible] 63

[illegible] $t_{0.025,20} = 2.086$ as $n = 23$ [illegible]

For β_1: $\hat{\beta}_1 \pm t_{0.025,20} SE(\hat{\beta}_1)$ [illegible]

$= 0.7 \pm 0.21 = (0.49, 0.91)$

and for β_2: $\hat{\beta}_2 \pm t_{0.025,20} SE(\hat{\beta}_2)$ [illegible] 2.086 × [illegible]

$= 0.2 \pm 0.21 = (-0.01, 0.41)$

Again, 95% confidence intervals for σ_u would be

$$P\left[(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,(n-3)}} \le \sigma_u^2 \le (n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,(n-3)}}\right] = 1-\alpha$$

when $\alpha = 0.05$, $n = 23$, $\hat{\sigma}_u^2 = 0.07$

$$P\left[20 \times \frac{0.07}{\chi^2_{0.025,20}} \le \sigma_u^2 \le 20 \times \frac{0.07}{\chi^2_{0.975,20}}\right] = 1 - 0.05 = 0.95$$

or $P\left[\frac{1.4}{34.170} \le \sigma_u^2 \le \frac{1.4}{9.591}\right] = 0.95$

or $P\left[0.041 \le \sigma_u^2 \le 0.146\right] = 0.95$

95% confidence intervals for σ_u are 0.041 and 0.146

To test the null hypothesis $H_0: \beta_1$ [illegible] the alternative $H_1: \beta_1 \neq 1$, the appropriate test statistic under the $H_0: \beta_1$ [illegible]

$$t\,(\text{observed}) = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} \sim t_{n-3}$$

$$\therefore\ t\,(\text{observed}) = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{0.7 - 1}{0.102} = \text{[illegible]}$$

Now $H_0: \beta_1$ [illegible] will be accepted if [illegible] and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,\,n-3} = t_{0.025,\,20} = 2.086$

Here we see that t (observed) $= -2.941$ does not lie in the interval [illegible] and [illegible] null hypothesis $H_0: \beta_1 = 1.0$ is rejected at 5% level of significance.

Again to test the null hypothesis $H_0: \beta_2$ [illegible] against the alternative H_1 [illegible] the appropriate test statistic under $H_0: \beta_2$ [illegible] would be

$$t\,(\text{observed}) = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} \sim t_{n-3}$$

$$\text{Here } t\,(\text{observed}) = \frac{0.2}{0.102} = 1.960$$

Now, $H_0: \beta_2 = 0$ will be accepted if for the given sample [illegible] observed [illegible] $\leq t_{\alpha/2,\,n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,\,n-3} = t_{0.025,\,20} = 2.086$

Here we see that t (observed) $= 1.960$ lies in the interval -2.086 and 2.086 and hence $H_0: \beta_2 = 0$ is accepted at 5% level of significance.

3.12. Prediction / Forecasting in the Multiple (Three-Variable) Regression Model

The formulas for prediction in multiple regression are similar to those in the case of simple (two variable linear regression) regression, except that to compute the standard error of the predicted value we need the variances and covariances of all regression parameters. Here we will present the expression for the standard error in the case of two explanatory variables.

Let the estimated regression equation be,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Now consider the prediction of the value Y_0 of Y given values X_{10} of X_1 and X_{20} of X_2, respectively. These could be values at some future date.

Then we have $Y_0 = \beta_0 + \beta_1 X_{10} + \beta_2 X_{20} + u_0$

and $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20}$

The prediction of Y [illegible]

[illegible]

where [illegible]

[illegible]

Thus, the prediction [illegible]

The variance of the prediction [illegible]

where [illegible]

var [illegible]

[illegible]

where var($\hat{\beta}$ [illegible]

and cov($\hat{\beta}$, $\hat{\beta}$ [illegible]

when [illegible]

When σ_u is not known, it is replaced by its unbiased estimator [illegible] where n = sample size. Thus, [illegible] confidence interval for the prediction would be $\hat{Y}_0 \pm t$ [illegible] SE($\hat{Y}_0$) where SE [illegible]

Example 3.11 Following Example [illegible]

(i) Using the estimated regression $\hat{Y}$ [illegible] Find the prediction of Y for X_2 = [illegible] and X_3 = [illegible]

(ii) Using the results of var($\hat{\beta}$ [illegible] and $\hat{\sigma}$ [illegible] derived in Example 3.8, estimate the variance of the prediction error and the standard error of the prediction error.

(iii) Find 95% confidence interval for the prediction.

Solution

(i) The prediction of Y can be obtained from the estimated regression equation.

$\hat{Y} = 111.70$ [illegible]

Here we put $X_{20} = 10$ and X_{30} = [illegible] and get

$\hat{Y}_0 = 111.70$ [illegible]

The predicted value of Y is $\hat{Y}_0$ = [illegible] when [illegible] and [illegible] = 400.

(ii) From example 3.8 we obtain var($\hat{\beta}$ [illegible] var($\hat{\beta}$ [illegible] 0.000[illegible] cov($\hat{\beta}$, $\hat{\beta}$ [illegible]

[illegible]

[illegible]

[illegible]

[illegible]

var$(\hat{Y}_0) = 52.24\left(1 + \frac{1}{10}\right) + (10 - 6)^2 \times 6.55$

$+ 2(10 - 6)(1400 - 800) \times 0.0$ [illegible] $+ (1400 - 800)^2 \times 0.000134$

Here $X_{10} = 10$ and $X_{20} = 1400$, which are given.

[illegible]

[illegible]

$= 324.504$

var$(\hat{Y}_0) = 324.504$

and $SE(\hat{Y}_0) = \sqrt{\text{var}(\hat{Y}_0)}$ [illegible]

Variance of the prediction would be [illegible] and standard error of prediction would be 18.014

(ii) We know that $(1 - \alpha)$ 100% confidence interval for prediction would be

$\hat{Y}_0 \pm t_{\alpha/2,\, n-k}\, SE(\hat{Y}_0)$ when $\alpha = 0.05$ [illegible]

95% confidence interval for prediction would be

$\hat{Y}_0 \pm t_{0.025,7}\, SE(\hat{Y}_0)$

or $58.56 \pm 2.365 \times 18.014$ or 58.56 ± 42.60 or (15.96, 101.16)

95% confidence interval for prediction would be 15.96 and [illegible]

3.13. Regression Analysis in Presence of Qualitative (Dummy) Variables

3.13.1. Meaning

In Section 1.10 we mentioned four types of variables that one generally encounters in empirical analysis. These are: ratio scale, interval scale, ordinal scale and nominal scale. The types of variables used in earlier sections were essentially ratio scale. In many cases we deal with models that may involve not only ratio scale variables but also nominal scale variables. Such variables are known as **indicator variables, categorical variables, qualitative variables**, or, **dummy variables (Binary variables)**.

3.13.2. Nature of Dummy Variables

In regression analysis the dependent variable or regressand is frequently influenced not only by ratio scale variables (e.g., income, output, prices, costs, height, weight, temperature, etc.) but also by variables that are essentially qualitative or nominal scale, in nature, such as sex, race, colour, religion, nationality, geographical region, political upheavals and party affiliation. For example, holding all other factors constant, female workers are found to earn less than their male counterparts or non-white workers are

[illegible]

[illegible] Such models are called **Analysis of variance (ANOVA) models**.

8.13.1 Use of Dummy Variables

The dummy variables are used [illegible] in the [illegible]. Dummy variables can be used for several purposes. We are explaining some of the uses of dummy variables in applied [illegible] research.

(i) Dummy variables as proxies to Qualitative (Categorical) Factors

Dummy variables are commonly used as proxies for qualitative factors such as profession, religion, sex, region etc. For example, we consider a sample of [illegible] from two regions of the country: rural and urban and we want to estimate the demand for [illegible] as a function of income. It is known that town dwellers [illegible] than rural [illegible]. Thus, region is an important explanatory variable in this case. We may represent this factor by a dummy variable to which we shall arbitrarily assign the value 1 for a town dweller and 0 for a person living in a rural area. The demand function can be written as

$$D = \beta_0 + \beta_1 Y + \beta_2 Z + u$$

where Y = income and Z = dummy variable for region. We put Z = 1 for a town dweller and Z = 0 for a person living in rural area.

(ii) Dummy variables as proxies of Numerical Factors

Dummy variables may be used as proxies for quantitative factors when no observations on these factors are available or when it is inconvenient to do so. For example, suppose we want to measure saving function $S = f(Y)$ from a cross-section sample of consumers. Obviously, 'age' is an important explanatory factor of the consumption and savings patterns of a consumer, since people become more thrifty with age. Although 'age' is a quantitative factor we may approximate it by a dummy variable. We may divide the selected consumers in two age groups:

Group 1 : People of 20–40 years of age

Group 2 : People of 40 years and over

On the assumption that people become more thrifty as they grow old, the dummy variable for 'age' may be assigned the value 0, if the person belongs to group-1 and the value 1, if the person belongs to group-2.

The saving function can be written in the form

$$S_t = \beta_0 + \beta_1 Y_t + [illegible]$$

where S = saving, Y = income

D = dummy variable [illegible]

where [illegible] > 0

(iii) Dummy variables are used for measuring the shift of a function over time

A shift of a function implies that the constant [illegible] changes while other coefficients remain constant. Such shifts may be [illegible] introduction of a dummy variable in the function.

For example suppose that we have data on the consumption function for an economy [illegible] period 1910-1968. During this period the economy faced [illegible] World Wars (1914-1918 and [illegible] 1945) and a deep depression (1929-1933). The abnormal conditions prevailing in these years have caused a shift of the consumption function [illegible] rationing, various controls and other factors. To capture the shift we may use a dummy variable say Z which would assume the value 0 during the above 'abnormal' years and 1 in the other normal years. The consumption function takes the form

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 Z_t + u_t, (\beta_2 > 0)$$

where C = consumption, Y = income,

Z = Dummy variable for the shift of the function

For a normal year the estimated form of the consumption function would be

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 Y + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 Y$$

and for an abnormal period it would be

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 Y$$

If we plot these two functions we can clearly see the shift of the consumption function during the abnormal (War and depression) years

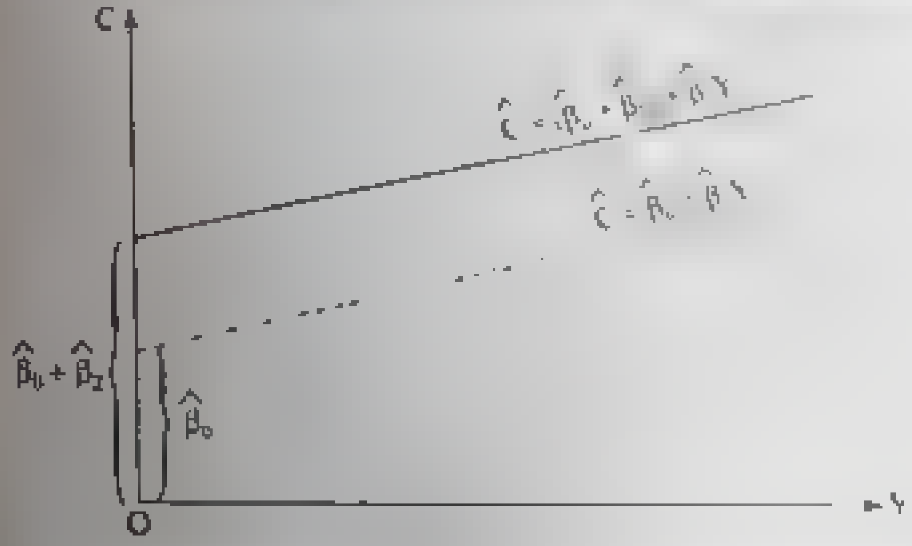


Fig. 3.1

The slope of the consumption function (Fig. 3.1) i.e. MPC is assumed to be the same both in normal and abnormal periods and hence the two regression lines are parallel (only intercept changes, slope remaining the same).

(iv) Dummy variables are used for measuring the change of parameters (slope) over time

It is usual that over the period of time [illegible] abnormal [illegible] not only the [illegible] intercept changes but also the [illegible] may well be expected to change [illegible] and propensities change over time [illegible] change in the parameters of a function may be captured by introducing appropriate dummy variables in the function.

We can write here the consumption function in the form

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 Z_{1t} + \beta_3 Z_{2t} + u_t$$

where C = consumption, Y = income

Z_1 = Dummy variable = 1 for abnormal years
= 0 for normal years

Z_2 = YZ_1 = Dummy variable = 0 for abnormal years (when Z_1 = 0)
= [illegible] for normal years (when Z_1 = [illegible])

Consequently, for a normal period the estimated consumption function would be given by

$\hat{C}$ = $\hat{\beta}_0$ + [illegible] Y, while for an abnormal year the estimated function would be

$$\hat{C} = \hat{\beta}_3 Y$$

In this case both slope and intercept of the function will change.

(v) Dummy variables are used as proxies for the dependent variable

In some cases, the dependent variable of a function may be a dummy variable. For example, suppose we want to measure the determinants of car ownership from a cross-section sample. Some people will have cars while others will not. Suppose that the determinants of the ownership depends on income and profession.

The functional relation can be written in the form

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 Z_i + u_i$$

where C = car owners or non-owners
= 1 for car owners
= 0 for non owner

Here C is taken as a dummy variable
Y = income
Z = a dummy variable for profession
= 1 if employed formally
= 0 if employed informally

It should be noted that if the dependent variable of a function is taken as a dummy variable, the disturbance term will be heteroscedastic and method of OLS will not be appropriate there.

(vi) Dummy variables are used for seasonal adjustments of time series

One of the most common use of dummy variables is in removing seasonal variations in time series. For example, if we have quarterly data on retail sales, we should adjust [illegible] milk per basis at [illegible] quarter [illegible] what [illegible] including the [illegible] and [illegible]

[illegible] the quarterly regression [illegible]

[illegible]

where Q_{1t} [illegible]

Q_{2t} [illegible]

Q_{3t} = 1 in the third quarter
0 in all other quarters

Dummy variable trap

It should be noted that we cannot introduce [illegible] with values [illegible] for the fourth quarter and zero [illegible] of the terms of sums of squares and sums of [illegible] including the quarterly dummies) would be zero [illegible] which is introduced with values equal to 1 in all [illegible] with the constant intercept β_0.

If we apply OLS to the above quarter's model [illegible] will give seasonal effect for each of the three [illegible] quarter [illegible] are zero and the seasonal effect for the fourth quarter [illegible] the constant intercept β_0.

In fact, when we introduce a large number of dummy variables into the model, we cannot obtain the OLS estimators of the parameters [illegible] matrix may be singular and $(X'X)^{-1}$ may not exist. This problem is called **Dummy variable trap.**

Some illustrative Examples .

Example 3.12. Consider the following model showing consumption expenditure by geographical region

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

where Y_i = Average consumption expenditure (₹) per person per 30 days in State i

$$D_{1i} = \begin{cases} = 1, \text{ if the State is in the Eastern region of India} \\ = 0, \text{ otherwise (i.e., in other region of the country} \end{cases}$$

$$D_{2i} = \begin{cases} 1, \text{ if the State is in the North Western region of the countr} \\ 0, \text{ otherwise (i.e., in other region of the country)} \end{cases}$$

Using data for 17 States of India in 2006-07 the following results are obtained by OLS method

$$\begin{array}{llll} \hat{Y}_i & = 1097.38 & - 241.04\, D_{1i} & - 30.09\, D_{2i} \\ SE & (103.31) & (133.37) & (129.50) \\ t & (10.62) & (-1.81) & (-0.23) \end{array}$$

The regression results show the mean per capita consumption expenditure [illegible] about ₹ [illegible] in the eastern region the per capita consumption [illegible] about ₹ [illegible] and that in the North-West Central region [illegible] lower [illegible] about ₹ [illegible]

Example 3.13. We consider a model to assess the determinants of literacy [illegible] across [illegible] States of India 2006-07.

The model takes the form

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

where Y = literacy rate (percent)

$$D_1 = \text{Gender} = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \text{Area of residence} = \begin{cases} 1 & \text{if urban} \\ 0 & \text{otherwise} \end{cases}$$

Using data of [illegible] States of India for 2006-07 the following results were obtained by OLS method:

$$\begin{array}{llll} \hat{Y}_i = & 75.82 & - 16.32\, D_{1i} & + 16.00\, D_{2i} \\ SE = & (1.82) & (2.10) & (2.10) \\ t = & (41.65) & (-7.77) & (7.62) \end{array}$$

In this regression model there are two dummy variables. The regression results show that the mean literacy rate is about 75.82 percent. Compared with this, the average literacy rate for females is lower by about 16.32 percent, for an actual average literacy rate of (75.82 − 16.32) = 59.50 percent.

By contrast for those who live in the urban area the mean literacy rate is higher by about 16 percent, for an actual average literacy rate of (75.82 + 16.00) = 91.82 percent.

Example 3.14. This example shows regression with a mixture of quantitative and qualitative regressors. We consider the following model:

Let Y = Average consumption expenditure (₹ per person for 30 days) in State

Let X = Average household size (the number of persons) in State

$$D_1 = \begin{cases} 1 & \text{if the State is in the Eastern region of India} \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if the State is in the North-West Central region of the country} \\ 0 & \text{otherwise} \end{cases}$$

The above equation is fitted with the help of the data on "Household consumer expenditure in India 2006-07" and obtained

$$\begin{array}{lllll} \hat{Y} = & 2454.72 & - 76.06\, D_1 & - 4.07\, D_2 & - 144.72\, X \\ SE & (503.[illegible]) & (145.7[illegible]) & (160.54) & (26.40) \\ t & (4.36) & (0.[illegible]) & (1.90) & ([illegible]) \\ & & & & R^2 = [illegible] \end{array}$$

These results suggest that other things remaining the same, as household size goes up by one person, on an average the per capita consumption expenditure goes down by about ₹ 144.72

14 A Brief Outline on Qualitative Response Regression Models

In all the regression models that we have [illegible] that the regressand, the [illegible] whereas the explanatory [illegible] of [illegible] or a mixture thereof.

[illegible] we may have to consider regression models in which the [illegible] qualitative in nature. The [illegible] response regression models [illegible] in various areas of social sciences and medical research.

For example, we like to study the labour force participation [illegible] whether [illegible] is either in the labour force or not [illegible]. Hence the response variable or regressand can take only two values, say [illegible] person is in the labour force and 0 if he/she is not. In other words, the regressand is a binary or **dichotomous variable**.

In qualitative regression models where the regressand [illegible] objective is to find the probability of something happening [illegible]. Hence qualitative response regression models are often known as probability [illegible].

There are four approaches to developing a probability model for a binary response variable where the regressand itself is qualitative in nature. These are:

1. The linear probability model (LPM)
2. The logit model
3. The probit model
4. The tobit model

Because of its comparative simplicity, and because it can be estimated by ordinary least square (OLS), we first consider the linear probability model (LPM).

The Linear Probability Model (LPM)

We consider a two variable regression model

$$Y_i = \alpha + \beta X_i + u_i \qquad (1) \text{ where}$$

X = family income, Y = a binary variable

i.e., $$Y = \begin{cases} = 1 & \text{if the family owns a house} \\ = 0 & \text{if it does not own a house} \end{cases}$$

Model (1) looks like a typical linear regression model but because the regressand is binary, it is called a **linear probability model (LPM)**. This is because the conditional expectation of Y_i given X_i, $E\left(Y_i/X_i\right)$, can be interpreted as the conditional probability that the event will occur given X_i, that is, $P(Y_i = 1 \mid X_i)$. Thus, in our example $E\left(Y_i/X_i\right)$ gives the probability of a family owning a house and whose income is the given amount X_i.

The justification of the name LPM for models like equation (1) can be seen as follows. Assuming $E(u_i) = 0$, as usual we obtain

$$E\left(Y_i/X_i\right) = \alpha + \beta X_i \qquad (2)$$

Now [illegible] probability that $Y_i = 1$ (that is, the event occurs) [illegible] and probability that [illegible] (that is, the event does not occur) [illegible] the variable [illegible] has the following probability distribution.

Y_i	probability
[illegible]	$1 - P_i$
[illegible]	P_i
[illegible]	[illegible]

This shows that Y_i follows a **Bernoulli probability distribution**.

Now, by definition of mathematical expectation

we obtain: $E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i$ ---- (3)

Now, comparing equation (2) with equation (3)

we can equate $E\left(\frac{Y_i}{X_i}\right) = \alpha + \beta X_i = P_i$ ---- (4)

that is, in fact, the conditional probability of Y_i. Since the probability P_i must lie between 0 and 1, we have the restriction.

$$0 \le E\left(\frac{Y_i}{X_i}\right) \le 1 \quad \text{———— (5)}$$

From the above explanation it would seem that OLS can be easily extended to binary dependent variable regression models. So we may assume that there is nothing new here. But this is not the case because the LPM poses several problems which are as follows.

(i) Non-Normality of the Disturbances u_i

Although OLS does not require the disturbances (u_i) to be normally distributed, we assumed them to be so distributed for the purpose of statistical inference. But the assumption of normality for u_i is not tenable for the LPMs because, like Y_i, the disturbances u_i also take only two values, that is, they also follow the Bernoulli distribution.

This can be seen clearly if we write equation (2) as $u_i = Y_i - \alpha - \beta X_i$ ---- (6)

The probability distribution of u_i is

	u_i	probability	
when $Y_i = 1$	$1 - \alpha - \beta X_i$	P_i	(7)
when $Y_i = 0$	$-\alpha - \beta X_i$	$1 - P_i$	

Obviously, u_i cannot be assumed to be normally distributed; they follow the Bernoulli distribution. But the non-fulfilment of the normality assumption may not be so critical as it appears because we know that the OLS point estimates still remain unbiased. Besides, as the sample size increases indefinitely, statistical theory shows that OLS estimators tend to be normally distributed generally. As a result, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

(ii) Heteroscedastic variances of the Disturbances

Even if $E(u_i) = 0$ and $\text{cov}(u_i, u_j) = 0$ for $i \neq j$ (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic. This is

[illegible]

[illegible] is not homoscedastic.

[illegible] but not efficient [illegible] they are not [illegible] methods of handling the heteroscedasticity [illegible]

depends on $E(Y_i/X_i)$, one way to resolve the heteroscedasticity problem [illegible] transform the model [illegible] by dividing it through by [illegible]

[illegible]

i.e. $$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_1}{\sqrt{w_i}} + \beta_2\frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \qquad (9)$$

We can now apply OLS in this model, called **Weighted Least Square (WLS)** [illegible] with w_i serving as weights.

In theory, what we have just described is fine. But in practice the true E [illegible] is unknown, hence the weights w_i are unknown. To estimate w_i we can use the following two-step procedure.

Step 1 : We can run the OLS on regression [illegible] despite the heteroscedasticity problem and obtain

$\hat{Y}$ = estimate of true $E\left(Y_i/X_i\right)$. Then obtain $\hat{w}_i = \hat{Y}_i(1-\hat{Y}_i)$, the estimate of w_i.

Step 2 We can use the estimated w_i to transform the data shown in equation (9) and estimate the transformed equation by OLS (i.e. weighted least squares).

(III) Non fulfilment of $0 \le E\left(Y_i/X_i\right) \le 1$

Since $E\left(Y_i/X\right)$ in the LPM measures the conditional probability of the event Y occurring, given X it must necessarily lie between 0 and 1. Although this is true *a priori* there is no guarantee that $\hat{Y}_i$, the estimators of $E\left(Y_i/X_i\right)$, will necessarily fulfil this restriction, and this is the real problem with the OLS estimation of the LPM. This happens because OLS does not take into account the restriction that $0 \le E(Y_i) \le 1$. There are the mean [illegible] fulfill this restriction [illegible]

estimate the LPM by the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1. If some are less than 0 (that is, negative), $\hat{Y}_i$ is assumed to be zero for those cases; if they are greater than 1, they are assumed to be 1. The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities $\hat{Y}_i$ will lie between 0 and 1. [illegible] estimated probabilities will indeed lie between the logical limits 0 and 1.

(iv) Questionable value of R^2 as a Measure of Goodness of Fit

The conventionally computed R^2 is of limited value in the dichotomous response models. Corresponding to a given X, Y is either 0 or 1. Therefore, all the Y values will lie either along the X axis or along the line corresponding to 1. [illegible] As a result, the conventionally computed R^2 is likely to be much lower than 1. In most practical applications the R^2 ranges between [illegible] R^2 [illegible] exceeds 0.8 when the predicted Y values are close to either 0 or 1.

Example 3.15 The following table (Table 3.8) gives invented data on home ownership Y (1 = owns a house, 0 = does not own a house) and family income X (thousands of dollars) for 40 families.

Table 3.8

Family	Y	X	Family	Y	X	Family	Y	X
1	0	8	15	0	6	29	0	11
2	1	16	16	1	19	30	0	10
3	1	18	17	1	16	31	1	17
4	0	11	18	0	10	32	0	13
5	0	12	19	0	8	33	1	21
6	1	19	20	1	18	34	1	20
7	1	20	21	1	22	35	0	11
8	0	13	22	1	16	36	0	8
9	0	9	23	0	12	37	1	17
10	0	10	24	0	11	38	1	16
11	1	17	25	1	16	39	0	7
12	1	18	26	0	11	40	1	17
13	0	14	27	1	20			
14	1	20	28	1	18			

From these data the estimated LPM (by OLS method) is given below

$$\hat{Y}_i = -0.9457 + 0.1021 X_i$$

$$SE \quad (0.1228) \quad (0.0082)$$

$$t = \quad [illegible] \quad (12.4512),\ R^2 = 0.8048$$

From the estimated LPM the intercept of −0.9457 gives the "probability" that a family with zero income will own a house. Since this value is negative and since probability

[illegible]

$$\hat{Y}_i(X = 12) = -0.9457 + 12 \times 0.1021 = 0.2795$$

This is the estimated probability that a [illegible] in particular. We [illegible]

We can note the probabilities [illegible] estimated probabilities will be negative and some will exceed 1.

For instance, when $X = 8$, $(\hat{Y}_i | X = 8) = -0.9457 + 8 \times 0.102$ [illegible]

Similarly, when $X = 20$, $(\hat{Y}_i | X = 20) = -0.9457 + 20 \times 0.102$ [illegible]

This means that although [illegible] need not be necessarily positive or less than 1. Thus [illegible] linear probability model when the dependent variable [illegible] even if the estimated Y are all positive and [illegible] problem of heteroscedasticity [illegible] model.

EXERCISE

1. In a multiple linear regression model, three variable [illegible] $i = 1, 2, \ldots, n$, why do we insert the random disturbance term [illegible]
2. State the assumptions about u_i of a classical linear regression model [illegible] model assumes the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2$ [illegible]
3. In a three variable linear regression model of the form [illegible] $i = 1, 2, \ldots, n$, how can you estimate the regression parameters [illegible]
4. Describe briefly the method of least squares used in estimating the regression parameters relating to a three variable linear regression model.
5. State and prove the properties of the least squares estimators relating to a three variable classical linear regression model (CLRM).
6. Show that in a three variable classical linear regression model [illegible] u_i, $i = 1, 2, \ldots, n$, the estimated parameters coefficients are unbiased.
7. In a CLRM, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, [illegible] express the estimated regression coefficients β_1 and β_2 in terms of variances and coefficients of correlations.
8. Determine the variances and covariances of the regression parameters in the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$.
9. State and prove GAUSS-MARKOV THEOREM in terms of a three variable linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$.

10. What is meant by the [illegible] regression [illegible]

[illegible]

11. How can you determine [illegible] $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$, $i = 1, 2, \ldots, n$?

12. In terms of a three variable linear regression model, show that [illegible] unbiased estimator [illegible]

13. What is a measure [illegible] three variable linear regression model [illegible] find

[illegible]

14. Show that in a three variable linear regression model [illegible] unbiased estimator of [illegible]

15. Show that the least squares estimators of the model [illegible] that

(i) [illegible]

(ii) [illegible]

where [illegible]

[illegible]

[illegible]

16. Show that MSE of [illegible] in a three variable linear regression model is not an unbiased estimator of σ_e^2 [illegible] unbiased estimator [illegible]

17. Describe the testing procedure of the significance of [illegible] parameters of the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$ [illegible]

18. What is meant by goodness of fit of a three variable linear regression model?

19. What is meant by multiple coefficient of determination? Derive the formula of multiple coefficient of determination in terms of a three variable linear regression model.

20. In terms of a [illegible] sum [illegible] squares.

21. In terms of a linear regression model [illegible]

show that [illegible]

where [illegible]

22. In terms of a three variable linear regression model [illegible] determination R^2 [illegible]

where [illegible]

23. Define multiple coefficient of determination R^2 [illegible] relation between R^2 and $\bar{R}^2$ when there are [illegible] model.

24. What is adjusted R^2? Why is it used [illegible]

25. Define partial correlation coefficients and show [illegible] in a three variable linear regression model.

26. Establish the relation among multiple [illegible] correlation coefficients and partial correlation [illegible] three variable linear regression model.

27. How can you formally write the regression model [illegible] $+ \beta_3 X_{3i} + u_i$ where [illegible] $2, \ldots, n$) satisfies all the properties of CLRM.

28. How can you use the analysis of variance in a three variable linear regression model?

29. What is the meaning of the term "predictor" [illegible] in a three variable linear regression model? [illegible] problem in this regard.

30. Establish the relation between R^2 and F in terms of a three variable linear regression model. What would be the value of F when $R^2 = 0$?

31. What is an ANOVA Table? How can you construct an ANOVA table in terms of R^2?

32. The Cobb-Douglas production function in its stochastic form is given by

$Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$ where Y = output, X_i = labour input, [illegible] stochastic disturbance term, β_1 = constant technological parameter. How [illegible] estimate the regression parameters by applying OLS method? [illegible] for the function?

33. What do you mean by indicator variables, categorical variables, qualitative variables, dummy variables, binary variables? Give an example.

34. What do you mean by dummy variables? How can you incorporate these variables in the regression model? What is meant by dummy variable trap?

35. What do you mean by dummy variables? Explain some of the uses of dummy variables in applied economic research.

36. What are the dummy variables? Construct a model where dummy variables are used as proxies for the dependent variable.

37. What are the dummy variables? Construct a model where dummy variables are used as proxies of numerical factors.

38. What are the dummy variables? Construct a model where dummy variables are used in measuring the shift of a function over time.

39. What do you mean by qualitative response regression models? Give a brief note on qualitative regression models and where they are applied [illegible].

40. The following results were obtained from [illegible] sets of observations on [illegible] [illegible] Estimate the regression of [illegible] and [illegible] [illegible] the results [illegible] that the coefficient of [illegible].

41. Consider the following regression model in deviation form [illegible] sample data [illegible]

$\Sigma y_i x_i = [illegible]$

(i) Compute OLS estimates of β_1, β_2 and R^2

(ii) Test the hypothesis $H_0: \beta_2 = 0$ against $\beta_2 \neq 0$

(iii) Test the hypothesis $H_0: \beta_1 = \beta_2 = 0$ against $H_1: \beta_1 \neq 0$ or $\beta_2 \neq 0$

(iv) Test the hypothesis $H_0: \beta_1 = 2\beta_2$ against $H_1: \beta_1 \neq 2\beta_2$

42. A production function model is specified as $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where Y = log output, X_1 = log labour input and X_2 = log capital input. The data refer to a sample of [illegible] firms and observations are measured as deviations from the sample means:

$\Sigma x_1^2 = [illegible]$, $\Sigma x_2^2 = [illegible]$, $\Sigma x_1 y = [illegible]$, $\Sigma x_2 y = [illegible]$, $\Sigma x_1 x_2 = [illegible]$, $\Sigma y^2 = [illegible]$

(i) Estimate β_1, β_2 and their standard errors

(ii) Find R^2 and adjusted R^2

(iii) Test the hypothesis that $\beta_1 = \beta_2$

(iv) Suppose now that you wish to impose the a priori restriction that $\beta_1 + \beta_2 = 1$. What is the least squares estimate of β_1 and its standard error? What is the value of R^2 in this case? Compare these results with those obtained in [illegible] and comment.

43. The following table shows 10 sets of values of three variables Y (dependent variable), X_1 and X_2 (two independent variables):

Y	3.5	4	[illegible]	6	[illegible]	9	8	[illegible]	2	4
X_1	5	20	30	42	50	54	65	72	85	90
X_2	16	13	10	7	7	5	4	3	3.5	2

(i) Consider a model of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. Find the least squares regression equation of Y on X_1 and X_2.

(ii) Compute the coefficient of multiple determination and the standard errors of the estimated parameters and conduct tests of significance.

(iii) Construct 95 percent confidence intervals for the population parameters and [illegible]

(iv) Find the explained and unexplained variation in Y.

[illegible]

[illegible]

(i) [illegible]

(ii) [illegible]

(iii) [illegible] coefficients.

[illegible] on the following [illegible] estimates [illegible]

Firms	A	B	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

(ii) Use the multiple [illegible]

(v) Construct 95% confidence interval [illegible]

The following table shows the value of imports [illegible] measured in arbitrary units and the [illegible] for the twelve-year period for a certain country.

2008	2009	2010	2011	2012	2013	2014	2015	2016	[illegible]	[illegible]	[illegible]
[illegible]	46	[illegible]	37	[illegible]	44	[illegible]	46	[illegible]	[illegible]	[illegible]	[illegible]
120	[illegible]	[illegible]	241	305	258	354	321	370	375	389	385
[illegible]	47	110	140	120	140	145	150	140	135	155	152

Estimate the import function $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

(i) What is the economic meaning of your estimates?

(ii) Construct tests of significance for the [illegible] estimates at 5% and [illegible] level of significance.

(iv) Compare R^2 and adjusted R^2.

The following table includes the output [illegible] the labour input [illegible] and capital input [illegible] of 15 firms of the chemical industry.

Firms	1	2	3	4	5	6	7	8
Y ('000 tons)	60	120	190	250	300	360	380	430
L (hours)	1100	1200	1430	1500	1520	[illegible]	[illegible]	[illegible]
K (machine hours)	300	400	420	400	510	590	600	630

Firms	9	10	11	12	13	14	15
Y ('000 tons)	440	490	500	520	540	410	350
L (hours)	1800	1750	1950	1960	1830	1900	1500
K (machine hours)	610	630	850	900	980	900	800

(i) Fit a Cobb-Douglas production function to the above data $Y = b_0 L^{b_1} K^{b_2} e^{u}$

(ii) Construct appropriate tests of significance of the parameter estimates at 1% and 5% levels of significance.

(iii) What are the marginal and average productivities of the factors L and K?

(iv) What do your results suggest regarding the returns to scale?

47. The following table shows the price index of durables, the average income of the household and expenditure on durables of a typical household of a country:

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018
Expenditure on durables (Y in ₹)	[illegible]	[illegible]	[illegible]	115	[illegible]	40	80	[illegible]	95
Household income (X_1 ₹)	833	2000	[illegible]	2040	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Price index (X_2)	100	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

(i) Fit a regression line to the function $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

(ii) Test your results by using the Analysis of variance table.

48. The following table shows the consumption of tobacco manufactures, consumer's income and the price of tobacco manufactures for France during 1950–59.

Year	Consumption (million tons) D	Income (million francs) Y	Price of tobacco (francs per kg) P
1950	59,190	76,200	23.56
1951	65,450	91,700	24.44
1952	62,360	106,700	32.07
1953	64,700	111,600	32.46
1954	67,400	119,000	31.15
1955	64,440	129,200	34.14
1956	68,000	143,400	35.30
1957	72,400	159,600	38.70
1958	75,710	180,000	39.63
1959	70,680	193,000	46.68

(i) Fit a linear regression $D = \beta_0 + \beta_1 P + \beta_2 Y + u$ and a non-linear function of the constant elasticity type $D = \beta_0 P^{\beta_1} Y^{\beta_2} u$

(ii) Conduct tests of significance using the analysis of variance table.

(iii) Compare the price and income elasticities of the two functions.

49. In a multiple regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ explain how you would test the joint hypothesis $\beta_1 = \beta_2$ and $\beta_2 = 1$.

50. Consider the following regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where $u_i \sim N(0, \sigma_u^2)$. The following data set are given below:

Y	4	7	3	9	[illegible]
X_1	2	3	1	5	9
X_2	5	3	2	7	

(i) Estimate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$

(ii) Find out var($\hat{\beta}_0$), var($\hat{\beta}_1$), var($\hat{\beta}_2$) and cov($\hat{\beta}_1, \hat{\beta}_2$)?

(i) Find R^2 and [illegible]

(ii) Write the regression [illegible]

(iii) Test β_0, β_1 [illegible]

(iv) [illegible] β_0, β [illegible]

(v) [illegible] β_1 [illegible]

(vi) Test [illegible]

(ix) Construct 95% and [illegible]

(x) Find out the point predictor [illegible] and [illegible]

(xi) Construct 95% confidence [illegible]

(xii) Test whether the effect [illegible]

50. Eight students made the following [illegible] certain subject. Fit the linear regression [illegible] score (X_1) and test score (X_2).

Students	1	2	3	4	5	6	7	8
Pretest score (X_1)	43	[illegible]	[illegible]	28	[illegible]	[illegible]	[illegible]	[illegible]
Test score (X_2)	[illegible]	29	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Final score (Y)	66	34	35	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Assume a linear regression equation of the form

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, [illegible]

(i) Find $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$

(ii) Find $var(\hat{\beta}_0)$, $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$ and $cov(\hat{\beta}_1, \hat{\beta}_2)$

(iii) Find R^2 and adjusted R^2

(iv) Write the regression results in the summary form

(v) Test whether β_0, β_1 and β_2 are significant or not at 5% level of significance

(vi) Find out the point predictor of Y when X_1 = 44 and X_2 = [illegible]

(vii) Construct 95% confidence interval of σ_u^2

(viii) Construct 95% confidence interval of the prediction.

51. Following exercise- 50, test the regression results by using Analysis of variance table

52. Following exercise- 5[illegible], test the regression results by using Analysis of variance table

53. The following table shows monthly income (X_1), monthly savings (Y) and age (X_2) of 10 persons.

Y	2000	8000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	40,000
X_1	[illegible]	20,000	[illegible]	30,000	40,000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
X_2	22	25	30	35	41	43	46	50	57	60

Assuming a linear regression [illegible]

and assuming [illegible] = 0 if [illegible] 40 years of age
[illegible] 40 years of age

(i) Find β_0, β_1 and β_2

(ii) Find R^2 and adjusted R^2

(iii) Write the regression results in the summary form

(iv) Find whether β_0, β_1 and β_2 are significant or not at 5% level of significance

(v) Estimate 95% confidence interval of [illegible]

55. The following regression was estimated from [illegible] quarterly data [illegible] in parentheses

[illegible]

(3.7) (0.27) (3.37) (5.40) (3.17)

where X_{it} = 1 in the ith quarter and 0 otherwise. Explain the implied pattern of seasonal variation and interpret the result.

56. You are given the following regression results

$\hat{Y}_t = 16.899 - 2978.5X_{2t}$ $R^2 = 0.6149$

(8.5152) (-4.7280)

$\hat{Y}_t = 9734.2 - 3782.2X_{2t} + 2815X_{3t}$ $R^2 = 0.7706$

(3.3705) (-6.6070) (2.9712)

Can you find out the sample size underlying these results ?

Hint: Recall the relationship among R^2, F and t values.

57. From the data for 46 States in the United States for a given year the following regression results were obtained

$$\log \hat{C} = 4.30 - 1.34 \log P + 0.17 \log Y$$

$$SE \quad (0.91) \quad (0.32) \quad (0.20) \qquad \bar{R}^2 = 0.27$$

where C = units of consumption of a commodity per year

P = real price per unit of the commodity

Y = per capita real disposable income

(i) What is the elasticity of demand for the commodity with respect to price ? Is it statistically significant ? If so, is it statistically different from 1 ?

(ii) What is the income elasticity of demand for the commodity ? Is it statistically significant ?

(iii) How would you retrieve R^2 from $\bar{R}^2$ given above ?

58. From a sample of 209 firms the following regression results were obtained

$\log(\text{salary}) = 4.32 + 0.280 \log(\text{sales}) + 0.0174\ \text{roe} + 0.00024\ \text{ros}$ $R^2 = 0.283$

SE (0.32) (0.035) (0.0041) (0.00054)

where salary = salary of CEO

sales = annual firm sales

roe = return on equity in percent

ros = return on firm's stock

Interpret the regression taking [illegible] have about the signs of the various coefficients.

(ii) Which of the coefficients are [illegible]

(iii) What is the overall significance [illegible]

(iv) Can you interpret the [illegible] or why not?

59. Consider the following wage-determination [illegible] period 1950-1969

$$\hat{W}_t = 8.582 + 0.364 PF_t + 0.004 PF_{t-1} \; [illegible]$$

$$(1.129) \quad (0.080) \quad (0.072) \quad [illegible]$$

where W = wages and salaries per employee

PF = price of final output at [illegible]

U = unemployment in Great Britain as a percentage [illegible] employees in Great Britain

t = time

(i) Interpret the regression equation.

(ii) Are the estimated coefficients individually significant?

(iii) What is the rationale for the introduction of PF?

(iv) How would you compute elasticity of wages and salaries per employee [illegible] to unemployment rate U ?

60. Consider the following data set

Y	[illegible]	3	8
X_2	1	2	3
X_3	2	1	−3

Based on these data, estimate the following regressions

$$Y_i = \alpha_0 + \alpha_1 X_{2i} + u_{1i} \qquad (1)$$

$$Y_i = \lambda_0 + \lambda_2 X_{3i} + u_{2i} \qquad (2)$$

$$Y_i = \beta_0 + \beta_1 X_{2i} + \beta_2 X_{3i} + u_i \qquad (3)$$

Estimate the regression coefficients in each case.

(ii) Is $\alpha_1 = \beta_1$? Why or why not ?

(iii) Is $\lambda_2 = \beta_2$? Why or why not ?

What important conclusion do you draw from this exercise.

61. From the following data estimate the partial regression coefficients, their standard errors and the adjusted and unadjusted R^2 values

$\bar{Y} = 367.693$, $\bar{X}_2 = 402.760$, $\bar{X}_3 = 8.0$

$\Sigma(Y_i - \bar{Y})^2 = 66042.269$ $\Sigma(X_{2i} - \bar{X}_2)^2 = 84855.096$

$\Sigma(X_2 - \bar{X}_2)^2 =$ [illegible] $\Sigma(Y - \bar{Y})(X_2 - \bar{X}_2) = 74778.346$

$\Sigma(X_3 - \bar{X}_3)^2 = 4250.900$ $\Sigma(Y - \bar{Y})(X_3 - \bar{X}_3) = 4796.00$ $n = 15$

62. Is it possible to obtain the following from a set of data ?

(i) $r_{23} = 0.9$, $r_{12} = -0.2$, $r_{13} = 0.8$

(ii) $r_{12} = 0.6$, $r_{23} = -0.9$, $r_{31} = -0.5$

(iii) $r_{21} = 0.01$, $r_{13} = 0.66$, $r_{23} = -0.7$

63. The regressed child mortality (CM) on percapita GNP (PGNP) and the female literacy rate (FLR) for a sample of 64 countries is given below

$$\widehat{CM}_i = 263.6416 - 0.0056\, PGNP_i - 2.2316\, FLR_i$$

$$SE \quad (11.5932) \qquad (0.0019) \qquad (0.2099)$$

$$R^2 = 0.7077, \quad \bar{R}^2 = 0.6981$$

(i) Interpret the regression results.

(ii) What about the statistical significance of the observed results ?

(iii) Is the coefficient of PGNP of −0.0056 statistically significant ?

(iv) Is the coefficient of FLR of −2.2316 statistically significant ?

(v) Are both the coefficients statistically significant jointly ?

4

Violations of Classical Assumptions— The Problems of Heteroscedasticity, Autocorrelation and Multicollinearity

4.1 Introduction

Let us consider a two variable linear regression [illegible]

u_i

If the model is a classical linear regression model [illegible] assumptions:

(i) u_i is a random variable

(ii) $E(u_i) = 0$ for each i

(iii) $E(u_i^2) = \sigma_u^2$ (or σ^2 constant)

(iv) $Cov(u_i, u_j) = E(u_i u_j) = 0$ for $i \neq j$

(v) X_i is non-stochastic or non-random.

We now put special consideration of assumptions (iii) and [illegible] $E(u_i^2) = \sigma^2$ means that the variance of the disturbance [illegible] the probability distribution of the disturbance term [illegible] feature of homogeneity of variance (or constant variance) [illegible] It may be the case, however, that all of the disturbance terms do not have the same variance. This condition of non-constant variance or non-homogeneity of variance is known as *heteroscedasticity*. Thus, we say that u's are heteroscedastic when

var $(u_i) \neq \sigma_u^2$ (or σ^2) [a constant value] but var $(u_i) = \sigma_{u_i}^2$ [illegible] (a value that varies)

Also we should not assume that each disturbance term has the same expected value equal to zero, i.e. $E(u_i) = 0$

If all the disturbance terms have expected value zero and same variance σ^2, then we say that all the disturbance terms are identically distributed.

If (iv) is also satisfied, it means that the different disturbance terms are independent of each other. So, when all (ii), (iii) and (iv) are satisfied, then we can say that the different disturbance terms are identically and independently distributed.

If the disturbance term varies from observation to observation, the different disturbance terms are not identically distributed.

Here var$(u_i) = E(u_i^2) = \sigma_{u_i}^2 \neq \sigma_{u_j}^2$ when $i \neq j$

This is the problem of *Heteroscedasticity*. The CLRM assumes that variance of the disturbance term is constant and if it is not constant, then the problem of

[illegible]

Autocorrelation is a special case of correlation. It refers to the relationship between the successive values of the same variable in different periods, while correlation refers to the relationship between two or more different variables. [illegible]

[illegible]

So, both the assumptions [illegible] of CLRM are violated then there is the problem of heteroscedasticity as well as the problem of autocorrelation.

4.2 Matrix Representation of Autocorrelation and Heteroscedasticity

Let us consider the dispersion matrix or variance-covariance matrix of the disturbance vector

[illegible]

[illegible]

If there is the problem of heteroscedasticity [illegible]

[illegible]

[illegible]

4.3 Consequences of the Problem of Autocorrelation and Heteroscedasticity

[illegible]

4.4 Consequences of the Problem of Heteroscedasticity

[illegible]

The OLS estimator of β is denoted by $\hat{\beta}$ [illegible]

[illegible]

[illegible] that the least square estimators are unbiased even under the [illegible]

[illegible]

[illegible]

4.5 Method for Estimating Regression Parameters in the Presence of the Problem of Heteroscedasticity

[illegible]

In the presence of the problem of heteroscedasticity we assume that $E(u_i^2) = \sigma_u^2 x_i$.

This means that the variance of the disturbance terms varies positively with the independent variable x_i.

If we apply the OLS method then we have $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$

Here $E(\hat{\beta}) = \beta$ because $E(u_i) = 0$

and $var(\hat{\beta}) = E(\hat{\beta} - E(\hat{\beta}))^2$

[illegible]

$$var(\hat{\beta}) = \frac{\sigma_u^2 \sum_{i=1}^{n} x_i^3}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \quad [\because E(u_i^2) = \sigma_u^2 x_i]$$

[illegible]

[illegible]

with $\hat{\beta}$ [illegible]

The WLS estimator of the parameter β denoted by β* has a lower variance than the variance of $\hat{\beta}$.

In this sense WLS method is more appropriate to a model which is subject to the problem of heteroscedasticity.

(Note: Heteroscedasticity may also be of the form [illegible])

4.6 Tests for Heteroscedasticity

There are three important tests for heteroscedasticity.

(i) Spearman's Rank Correlation test
(ii) Goldfeld and Quandt test
(iii) Glejser's test.

All these test criteria are based on the OLS residual terms.

Let us consider the following model $Y = \alpha + \beta X + u$ where [illegible]

[illegible] of the given observations on X and [illegible] and ignoring the [illegible]

[illegible]

[illegible]

[illegible] Spearman's Rank Correlation Test

[illegible] applicable for large or small samples.

[illegible] the rank correlation coefficient between [illegible]

[illegible] coefficient is computed by the formula

[illegible]

[illegible] difference between the ranks of corresponding pairs of X and [illegible] and [illegible] observations in the sample.

[illegible] the null hypothesis that the rank correlation coefficient [illegible] hypothesis that it is not equal to zero [illegible] alternative hypothesis H_1 [illegible]

The appropriate test statistic is then given by [illegible]

[illegible] degrees of freedom.

[illegible] of the given sample for the null hypothesis H_0 [illegible] will be accepted [illegible]

If the null hypothesis is accepted then there is no problem of [illegible]. But if it is rejected then there will be the problem of [illegible].

4.6.2 Goldfeld and Quandt Test

[illegible] applicable to large samples. For this test we have to consider the [illegible]

(i) [illegible] to order the observations according to the magnitude of the [illegible] variable [illegible]

(ii) [illegible] select an arbitrary number [illegible] and omit a certain number of central [illegible] from the analysis. The remaining [illegible] observations are then divided into two equal parts. Each part therefore will contain [illegible] observations. One part includes the small values of X while the other part [illegible] the large values of X.

(iii) [illegible] regression lines by OLS procedure to each part and obtain the residual [illegible] from each of them.

Let S_1 denote the sum of squared residuals from the sample of low values and S_2 denote the same from the sample of large values of [illegible]. Then we calculate

[illegible]

[illegible] test the null hypothesis

[illegible] against the alternative hypothesis

[illegible]

4.6.3 Glejser's Test

[illegible]

Now we have to test the null hypothesis [illegible] against the alternative hypothesis [illegible]

[illegible]

Example [illegible] A researcher using time series data for the period [illegible] estimated the following consumption function [illegible]. The following table gives the data used and the residual errors.

Year	1954	[illegible]	1956	1957	1958	1959	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Consumption (billions of $)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Income Y (billions of $)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	0.52	[illegible]	[illegible]	[illegible]	2.99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

Solution [illegible] Spearman's rank correlation [illegible]. The findings are shown in the following table: Table 4.3

[illegible]

[illegible] the null hypothesis that the rank correlation [illegible] against the alternative hypothesis that [illegible]

Then we have

[illegible]

[illegible] the null hypothesis will be accepted for the given sample at 5% level of significance.

[illegible]

Example 4.5 [illegible]

[illegible]

Solution [illegible]

[illegible]

Autocorrelation

[illegible]

[illegible] substitute u_{t-1} then we get,

[illegible]

[illegible] substitute u_{t-2} then we get,

[illegible]

substitute u_{t-3} then we get,

[illegible]

[illegible]

Thus [illegible]

[illegible] autoregressive [illegible]

[illegible] Mean, Variance and Covariance of the Autocorrelated Disturbance Variable

1. Mean of the autocorrelated u's

[illegible]

2. Variance of the autocorrelated u's

By definition [illegible]

[illegible]

(Follows from the [illegible])

[illegible]

3. Covariance of the autocorrelated u's

Since [illegible]

and [illegible]

Now, [illegible]

[illegible]

[illegible]

[illegible] and in general [illegible]

[illegible] on the basis of (1), (2) and (3), we find that when u_i's are [illegible]

[illegible] and [illegible] $\neq 0$

[illegible] Consequences of Autocorrelation

[illegible]

OLS estimators are unbiased

[illegible] simple regression model we get [illegible] where

and [illegible]

[illegible]

[illegible]

- The variances of OLS estimates are underestimated

[illegible]

$var(\hat{\beta}) = \sigma_u^2 / \sum x_i^2 = \frac{\sigma_u^2}{\sum x_i^2}$

[illegible]

However, with u's related with a first order autoregressive scheme we have

[illegible]

Therefore var($\hat{\beta}$) [illegible] Expanding the second term in the above expression we obtain

[illegible]

- [illegible] will be inefficient through [illegible]

The Predictions will be inefficient

[illegible]

Test for Autocorrelation

[illegible]

Durbin-Watson Test

[illegible] Watson have suggested a test which is applicable [illegible] However, the test is appropriate only for the first order autoregressive [illegible]

[illegible] we are interested in testing the null hypothesis [illegible] are not autocorrelated with a first order [illegible] are autocorrelated [illegible] the null hypothesis we use the Durbin-Watson statistic [illegible]

[illegible] where e [illegible] are the OLS residuals [illegible]

[illegible]

For very large values of n, the large samples $\sum$ [illegible], $\sum$ [illegible] and $\sum$ [illegible] are approximately equal.

Then d [illegible]

But [illegible] where $\hat{\rho}$ is the [illegible] estimator of ρ in the model [illegible]

[illegible] is estimated.

Since [illegible]

$0 \le d \le 4$

This shows that d lies between 0 and 4.

Firstly, if there is no autocorrelation $\hat{\rho} = 0$ and $d = 2$. Thus if from the sample we find that d (observed) $= d^* = 2$, we accept that there is no autocorrelation in the function.

Secondly, if $\hat{\rho} = +1$, $d = 0$ and we have perfect positive autocorrelation.

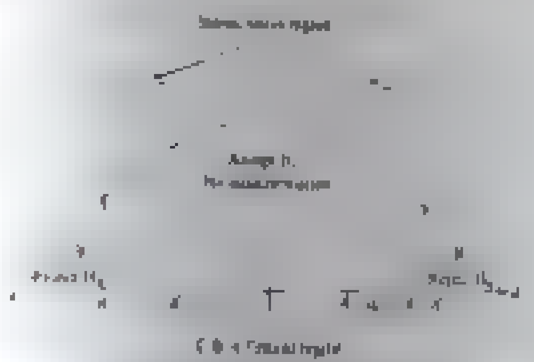
Thirdly, if $\hat{\rho} = -1$, $d = 4$ we have perfect negative autocorrelation. Therefore if $2 < d^* < 4$ there is some degree of negative autocorrelation, which is stronger if the value of d^* is higher.

If there is no problem of autocorrelation, $\hat{\rho}$ should be zero and $d = 2$.

Thus we have to test the null hypothesis $H_0: d = 2$ against the alternative $d \neq 2$.

Here the problem is that the exact sampling distribution of the statistic d is not known. What Durbin and Watson have done is to specify the upper and the lower limit of d.

Let d_U stand for the upper limit of d and d_L stand for the lower limit of d. This is shown in the diagram on the next page (Fig. 4.2).



[illegible] critical region

Fig. 4.2 The critical regions of d test

Thus with the help of d_L and d_U we have to determine whether autocorrelation exists or not. The values of d_L and d_U are available in the DW tables [illegible] of n and [illegible]

We calculate $d^* = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$ from the sample.

We may now consider the following cases:

1. If it is found that $d^* < d_L$ we reject the null hypothesis of no autocorrelation and accept that there is positive autocorrelation of first order.
2. If $d^* > 4 - d_L$ we reject the null hypothesis of no autocorrelation and accept that there is negative autocorrelation of the first order.
3. If $d_U < d^* < 4 - d_U$ we accept the null hypothesis of no autocorrelation.
4. If $d_L \le d^* \le d_U$ or if $4 - d_U \le d^* \le 4 - d_L$ the test is inconclusive.

Limitations of Durbin-Watson Test

There are some limitations of Durbin-Watson test.

1. There exist inconclusive regions and hence if the value of d^* lies between either d_L and d_U or between $4 - d_U$ and $4 - d_L$, then we cannot conclude whether autocorrelation exists or not.
2. This test method is appropriate only when the nature of the autocorrelation is of first order autoregressive type. But it is not appropriate when autocorrelation is of higher order and non linear type.
3. If there is any lagged variable as independent variable in the model, then Durbin-Watson statistic d is inappropriate in testing for autocorrelation.

[illegible] level of significance with [illegible] and [illegible] explanatory [illegible] and [illegible] 36. Here we see that [illegible] and hence [illegible]

[illegible] on the basis of Durbin-Watson test we [illegible] [illegible] problem of [illegible]

4.10.3 Von Neumann Ratio Method of Testing Autocorrelation

[illegible] the ratio of the variance of the first differences of any variable [illegible] variance of [illegible] is given by

$$[illegible]$$

The Von Neumann [illegible] is applicable for [illegible] which are random [illegible] where the [illegible] are not [illegible] in the case of the [illegible] [illegible] estimated from the OLS residuals [illegible] large samples [illegible] the Von Neumann [illegible]

$$VN = [illegible]$$

This test statistic is used for [illegible] the existence of autocorrelation.

This procedure is [illegible] applicable for testing the autocorrelation of the [illegible] especially if the sample is small [illegible]

4.11 Methods for Estimating Regression Parameters in the Presence of the Problem of Autocorrelation

When the autocorrelation is detected, the appropriate remedy or procedure [illegible] [illegible] in the original data is transforming [illegible] OLS is applied to the transformed data.

Let the model be given by [illegible]

where [illegible]

If we take a lagged form of the model [illegible] and multiply both sides by [illegible]

[illegible]

Now subtracting [illegible] from [illegible] we get

[illegible]

or, [illegible]

or, [illegible]

where [illegible] and $\alpha^* = [illegible]$

Here [illegible] satisfies all the properties of CLRM.

It should be noted that in transforming [illegible] observation shall be lost because of lagging and subtracting [illegible] from [illegible] we can apply OLS to the transformed relation [illegible] obtain α^* and β for our parameters [illegible]

[illegible] and [illegible] can be shown that

[illegible] because [illegible] is perfectly and linearly related to [illegible] Again

[illegible] all standard assumptions of [illegible] the estimates [illegible] and [illegible] our standard OLS formulae.

$$[illegible]$$

[illegible] where [illegible] the regression [illegible]

[illegible] to obtain [illegible]

[illegible] only [illegible] sample size is large [illegible]

[illegible] only if the value of [illegible] is known.

[illegible] the methods through which [illegible] can be estimated.

Method I : Using [illegible] information on ρ

[illegible] make [illegible] reasonable guess about the value of the [illegible] by using his knowledge or intuition about the [illegible] relationship [illegible] Now the transformed model becomes

[illegible]

[illegible]

[illegible] method we can estimate only [illegible] but [illegible] cannot be estimated.

[illegible] then [illegible] and β can be estimated at a time.

Method II : Estimation of ρ from the d-statistic.

From the Durbin-Watson test statistic we know that [illegible]

where [illegible]

Suppose that we calculate certain value of d statistic = [illegible] from sample data.

then [illegible] so that [illegible]

When [illegible] is estimated then we can estimate [illegible] and β from the model.

It should be noted that [illegible] will not be accurate if the sample size is small. This is [illegible] only for large samples.

[illegible]

Example 4.4 [illegible]

Solution [illegible]

[illegible] assumption [illegible] and the variable [illegible]

[illegible] In order to [illegible]

[illegible] after substitution [illegible] Table 4 [illegible]

As the level of significance for [illegible] the tabulated [illegible] Here we see that [illegible] and hence we reject the null hypothesis [illegible] autocorrelation [illegible] the alternative hypothesis [illegible] disturbance terms.

(b) Assuming a first order autoregressive scheme of the form [illegible] we may estimate the value of ρ given [illegible]

[illegible]

[illegible] regression model of the transformed form is

[illegible]

[illegible] be stated in terms of the original variables [illegible]

[illegible] Estimation by Levels versus First Differences

[illegible] we reject the hypothesis of zero serial correlation, what is the best [illegible]

[illegible] we may estimate a regression by transforming all the variables by [illegible] differencing, that is, regress [illegible] where [illegible] is the estimated [illegible] is subject to sampling errors, one other alternative that is [illegible] is to use a first difference equation. In fact a rough rule

[illegible]

[illegible] Multicollinearity—Meaning and Sources

[illegible]

[illegible]

[illegible] Consequences of Multicollinearity

[illegible]

[illegible]

If there is perfect multicollinearity [illegible]

and hence [illegible] and hence [illegible]

[illegible]

Hence $\hat{\beta} = \frac{0}{0}$ remains indeterminate if [illegible]

This means that OLS estimates of the parameters cannot be obtained if there is exact (perfect) multicollinearity in the regression model.

[illegible]

Thus [illegible] are the parameters are indeterminate.

[illegible] the standard errors of estimates become infinitely large [illegible] in the regression model [illegible] are perfectly correlated [illegible], the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be

[illegible]

Substituting [illegible] we get

[illegible] undefined or infinitely large

Similarly [illegible] undefined or infinitely large

[illegible]

[illegible] that $\hat{\beta}_1 - \beta_1$ is the sampling error of the estimator depends on the correlation coefficient between [illegible] and [illegible]

[illegible]

[illegible] shows that the sampling error of the estimator $\hat{\beta}$ depends on the [illegible] of the correlation coefficient between X and [illegible]

Hence the higher the degree of multicollinearity the higher will be the variance of the estimates of the parameters. In other words, if $r^2_{X_1X_2}$ or $|r_{X_1X_2}|$ is high, variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are also high. Hence the BLUE properties of the estimators no longer remain valid. This can be proved as follows.

Proof: We know that for a three variable linear regression model, variance-covariance matrix of $\hat{\beta}$ is given by, [illegible] (See Section 3.2)

[illegible]

Multicollinearity—Meaning and Sources

[illegible]

variables is equal to zero), [illegible] orthogonal [illegible]

[illegible]

variables due to the interdependence of more [illegible]

Multicollinearity is not a condition that [illegible] but rather a phenomenon inherent in [illegible] will affect seriously the parameter [illegible]

[illegible] difficult to [illegible] for example [illegible] dependent [illegible] in the same proportion, [illegible]

Consequences of Multicollinearity

[illegible]

Exact Multicollinearity and its Consequences

[illegible]

[illegible]

[illegible]

It should be noted that when [illegible] TOL = 0 and when [illegible] [illegible]

2. Wider confidence intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger [illegible]

Table 4.5

The effect of increasing collinearity on the 95% confidence interval for β_2

[illegible]

where VIF [illegible]

[illegible]

3. Insignificant t Ratios

[illegible]

4. A high R^2 but few significant t Ratios

[illegible]

5. Sensitivity of OLS Estimators and their Standard Errors to Small Changes in Data

[illegible]

Table 4.6 [illegible] Table 4.7 [illegible]

[illegible]

$\text{cov}(\hat\beta_1, \hat\beta_2) =$ [illegible] | $\text{cov}(\hat\beta_1, \hat\beta_2) =$ [illegible]

As a result of a slight change in the data we see that $\hat\beta_1$ which was statistically significant before at 10% level of significance, is no longer significant even at that level. It is also observed that [illegible] has [illegible] almost a three fold increase. All these changes may be attributable to increased multicollinearity as r_{12} has been increased from [illegible]. Accordingly, the standard errors of $\hat\beta_1$ and $\hat\beta_2$ have also increased between the two regressions, a usual symptom of collinearity.

4.5. Some Illustrative Examples

We can discuss some examples where the intercorrelations between the explanatory variables are high and study the consequences. Consider the model

[illegible]

[illegible]

Thus, only $(\beta_1 + 2\beta_2)$ would be estimable. We cannot get estimates of β_1 and β_2 separately. In this case we say that there is "perfect multicollinearity" because x_1 and x_2 are perfectly correlated (with $r_{12} = 1$). In actual practice, we encounter cases where r^2 is not exactly 1 but close to 1.

As an example, we consider the case where $\Sigma x_1^2 = 200$, $\Sigma x_1 y_1 = 350$, Σ [illegible], $\Sigma x_2 y_1 = 263$ and $\Sigma x_2^2 = 113$

[illegible]

[illegible] which is very high.

[illegible]

Thus the variance of $\hat\beta_1$ will be high if [illegible] σ_u is high, [illegible] is low, [illegible] is high. [illegible] is low and [illegible] is high we will not have the problem of high standard errors. On the other hand, even if [illegible] is low, the standard errors can be high if σ_u is high and [illegible] is low [illegible] there is not sufficient variation in [illegible].

What this suggests is that high values of [illegible] do not tell us anything about whether we have a multicollinearity problem or not.

Example [illegible]

Table 4.8

[illegible]

Assuming that [illegible] influenced by [illegible] the table [illegible] is given by

$E = \beta_0$ [illegible]

Where E = Expenditure on clothing

[illegible]

[illegible] following estimated regression results were obtained

[illegible]

Applying analysis of variance [illegible] test [illegible] we find

F (observed) = [illegible]

Here K = number of parameters including the constant intercept term [illegible]

and n = number of years (sample size) = [illegible]

From the table value we see that [illegible] and we reject the null hypothesis, accepting the alternative that there is a significant relationship between clothing expenditure and the explanatory variables.

However, all the explanatory variables are [illegible] as can be seen from the simple correlation coefficients

[illegible] = 0.973, [illegible]

[illegible] we compute the elementary regressions

[illegible]

R^2 [illegible]

[illegible] the first [illegible] the period under [illegible]. We then introduce the remaining explanatory variables gradually [illegible]

results are shown in the following table

Table 4.9

[illegible]

Note: [illegible] numbers in brackets are the standard errors of the estimates.

[illegible] explaining the [illegible] expenditure. The introduction of [illegible] improves slightly the overall R^2. [illegible] standard errors show that [illegible] does not affect the significance of [illegible] not give a satisfactory estimate for either [illegible] makes it impossible to obtain separate [illegible] However, the estimate [illegible] is not affected, despite the [illegible] Thus [illegible] may be considered as a superfluous [illegible]

[illegible] introducing [illegible] in the function we obtain a higher overall fit. [illegible] increased and all the parameter estimates have correct signs and are statistically [illegible] despite the high degree of [illegible] the standard errors [illegible]

[illegible]

4.16 Tests for Detecting Multicollinearity

[illegible]

1. High R^2 but few significant t ratios

[illegible]

2. High pair-wise correlations among regressors

[illegible]

[illegible]

3. High value of R^2 and low values of partial correlation coefficients

[illegible]

[illegible]

4. The test suggested by Farrar and Glauber

[illegible]

...with the population.

4.17 Solutions to the Problem of Multicollinearity

There are the commonly used methods to solve the problem of multicollinearity:

(1) Dropping of variables
(2) Using extraneous estimates
(3) Ridge Regression
(4) Using ratios or first differences
(5) Using principal components
(6) Getting more data

(1) Dropping of variables

[illegible]

[illegible] are unknown and hence [illegible] is to be replaced [illegible]

[illegible]

[illegible] according as [illegible]

unknown, we have to compute the value of [illegible]

then MSE [illegible] = var [illegible]

[illegible] then MSE [illegible] < var [illegible]

[illegible]

(b) Use of Extraneous Information

[illegible]

method. Let $\hat{\beta}$ be the OLS estimator of β. Then $\hat{\beta}$ is the [illegible] elasticity of demand from the cross section data.

[illegible]

Ridge Regression

[illegible] HOERL and KENNARD [illegible]

[illegible]

[illegible] are the estimators of β_i from the ridge regression, the β_i are the true values and k is the number of regressors.

[illegible] a function of the regression parameters β_i and the error variance [illegible] HOERL and KENNARD suggest trying different values of k [illegible] Some others have suggested obtaining initial estimates of [illegible] use the estimated k. This procedure can be iterated and we get

[illegible]

[illegible] using first differences

[illegible]

[illegible] it may be that the degree of correlation between [illegible] and [illegible] may be lower than the degree of correlation between [illegible] and [illegible]

The CLRM properties are

$E(u_t)=0,\ E(u_t^2)=\sigma_u^2,\ E(u_t\,u_{t-1})=0$

$E(\Delta u_t^2) = E(u_t - u_{t-1})^2$

$= E(u_t^2 - 2u_tu_{t-1} + u_{t-1}^2)$

[illegible]

[illegible] the first differences of the dependent variable [illegible] correlation between the disturbance terms [illegible] are not independent of each other. This is called the problem of [illegible]

[illegible] Principal Components

[illegible] principal [illegible]

[illegible]

[illegible] has the highest variance (subject to the normalization [illegible]

[illegible] are called the principal components [illegible] can be ordered [illegible]

[illegible] with highest variance is called the first principal component, [illegible] highest variance [illegible] called the second principal component and so on. [illegible] components have the following properties

$\mathrm{var}(Z_1) + \mathrm{var}(Z_2) + \dots + \mathrm{var}(Z_k) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \dots + \mathrm{var}(X_k)$

[illegible] which are uncorrelated, the Z's are orthogonal or uncorrelated [illegible] there is zero multicollinearity among the Z's.

[illegible] suggested [illegible] of regressing [illegible] X_k, we should [illegible] But this is not a solution [illegible] the multicollinearity problem. [illegible] and then substitute the values of [illegible] in terms of [illegible] we finally [illegible] the same answers as before. The fact that the [illegible] are uncorrelated does not mean [illegible] get better estimates of the coefficients in the original regression equation. [illegible] using the principal components only [illegible] we regress Y on a subset [illegible]

[illegible]

(e) Getting more data

It has been suggested that multicollinearity may be avoided [illegible] increasing the size of the sample [illegible] increasing the sample [illegible] sample size [illegible] sample that [illegible] the population [illegible]

EXERCISE

1. What is meant by autocorrelation ? How does it differ from simple [illegible] assumption of the CLRM will not hold if the problem of autocorrelation exists ?
2. What is the meaning of the term heteroscedasticity [illegible] ? What assumption of the CLRM will not hold if the problem of [illegible] exists in the model ?
3. By using matrix algebra how can you identify the problem of [illegible] the problem of heteroscedasticity ?
4. Explain the possible consequences of the problem of heteroscedasticity.
5. Show that the OLS estimators will be unbiased even if there is a problem of [illegible] scedasticity in the CLRM.
6. Show that in presence of heteroscedasticity the OLS estimators will be unbiased but BLUE property may not be satisfied.
7. How can you estimate the regression parameters in presence of heteroscedasticity ?
8. Explain the WLS method of estimating a regression model in the presence of [illegible] of heteroscedasticity.
9. Explain briefly the different testing procedures used for testing [illegible] heteroscedasticity.
10. Given $Y_i = \alpha + \beta X_i + u_i$ with $E(u_i^2) = [illegible]$. Prove that OLS estimators [illegible] possess greater variance than the OLS estimates of the transformed model [illegible]

[illegible]

(Hint : [illegible])

Here [illegible] OLS estimation of the transformed model.

11. Consider the model $Y_i = \alpha + \beta X_i^2 + u_i$ [illegible]

Assume that the variance [illegible] Show that [illegible] approximations [illegible] formula for the variance of [illegible]

12. Explain Durbin-Watson [illegible] testing the problem of autocorrelation. Find out the critical boundaries of Durbin-Watson test.

[illegible]

23. What is meant by multicollinearity? What are the [illegible] the multicollinearity in a regression model?

24. What is meant by multicollinearity? Explain the following [illegible] multicollinearity
a. Klein's rule for detecting multicollinearity
(b) The test suggested by FARRAR and GLAUBER

25. State and briefly explain the methods used for solving the problem of [illegible]

[illegible]

[illegible]

[illegible]

5

Specification Analysis

[illegible] Introduction

[illegible]

5.2 Diagnostic Tests Based on Least Squares Residuals

[illegible] problems with the [illegible]

We are [illegible] based on least squares residuals.

[illegible] regression [illegible] specified [illegible] including all the relevant variables [illegible] excluding all irrelevant [illegible] specified [illegible] specification error or model specification bias.

[illegible] supply from the model must be logically possible

[illegible]

The residuals remaining from the model must be purely random [illegible]

[illegible]

5.4 Types of Specification Errors

[illegible] model should satisfy the above listed criteria [illegible]

where C = total cost of production and Q = output [illegible]

[illegible]

Now assume that another researcher uses the following model

[illegible]

Now assume that another researcher formulates the following model

[illegible]

[illegible] bias, the bias here being the use of wrong functional form. In equation [illegible] Q appears linearly whereas in equation [illegible] Q appears log linearly

[illegible]

errors of measurement bias.

[illegible]

[illegible]

1. [illegible]
2. Inclusion of unnecessary (irrelevant) variable(s)
3. Adoption of the wrong functional form
4. Errors of measurement
5. Incorrect specification of the stochastic error term
6. Assumption that the error term is normally distributed

[illegible] model specification errors, [illegible] model mis-specification errors [illegible]

In the present book we will concentrate mainly on model specification errors [illegible]

Consequences of Model Specification Errors

[illegible]

Underfitting a Model (Omitting a Relevant Variable)

[illegible]

[illegible]

Example [illegible]

[illegible]

SE [illegible]

[illegible]

Assuming the above stated model as the [illegible] true model [illegible] variable from the model and examine its consequence. [illegible]

[illegible]

The regression results are

[illegible]

SE [illegible] $r^2 = $ [illegible]

[illegible]

[illegible] Inclusion of an Irrelevant Variable (Overfitting a Model)

[illegible]

Proof The proofs of some of the above statements are given below.

[illegible]

[illegible]

[illegible]

From the true model [illegible]

[illegible]

we have [illegible]

[illegible]

Note [illegible]

[illegible]

[illegible] Tests of Specification Errors

[illegible]

[illegible] Detecting the Presence of [illegible] Variables ([illegible] a Model)

[illegible]

[illegible] bottom up approach [illegible] data mining [illegible] regression fishing, data grubbing, data snooping [illegible]

5.6.2 Tests for Omitted Variables and Incorrect Functional Form

[illegible]

(i) Examination of Residuals

[illegible]

but [illegible] researcher fits the following quadratic cost function

[illegible]

and another researcher fits the following linear cost function

[illegible]

where [illegible] total cost [illegible] output

[illegible] is the true cost function then both the researchers [illegible] specification [illegible]

Let us consider an example.

Example 5.2. We have the following set of data on output [illegible] and total cost [illegible]

[illegible]

[illegible]

Table 5.1

Estimated residuals from the Linear, Quadratic and Cubic cost functions

[illegible]

Fig. 4 [illegible]

[illegible] the diagram Fig. [illegible] we see that [illegible] the residuals [illegible] effects.

(d) The Durbin Watson d Statistic

The Durbin Watson statistic d is usually computed by the formula

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} \qquad \text{(See Chapter 4)}$$

where e_t = Y − [illegible] = observed value of Y − estimated value of Y

We examine the computed (calculated) Durbin Watson Statistic d of the [illegible] models (linear, quadratic and [illegible]). We see that for the linear [illegible] d = [illegible], suggesting that there is positive correlation in the estimates. [illegible] for n = [illegible] and k = [illegible] (number of explanatory variables excluding the constant) 5% critical d values are d_L = 1.270 and d_U = [illegible]. Similarly the computed d [illegible] (ii) the quadratic cost function is [illegible] whereas the 5% critical d values are [illegible] d_L = 0.69 and d_U = [illegible] indicating indecision (inconclusive) [illegible]

[illegible]

the assumed model, since the [illegible] O.L.S. residuals.

[illegible]

[illegible] that the data are [illegible]

[illegible] Durbin Watson tables, if the computed d value is significant, then we can accept the [illegible] specification.

[illegible]

[illegible] for example the procedure [illegible] the quadratic [illegible] model.

(e) Ramsey's RESET Test

Ramsey has proposed a general test of specification error called RESET [illegible] we are explaining the [illegible] output example [illegible] cost function of the form

$$[illegible]$$

where [illegible] total cost and [illegible] output

[illegible]

[illegible]

[illegible]

where R^2_{new} = 0.9983, R^2_{old} = 0.8409

n = [illegible] number of [illegible] number of parameters in the new model = 4

[illegible] indicating that the model

Y_i = [illegible]

se (19.02) [illegible]

is mis-specified.

Of course, we have reached [illegible] the same conclusion on the basis of the visual examination of the residuals [illegible] Durbin-Watson d value.

However, one advantage [illegible] is that it is easy to apply [illegible] require one to specify what the alternative model is. But this is also its disadvantage because knowing that a model [illegible] does not help us necessarily in choosing a better alternative.

(iv) Lagrange Multiplier (LM) Test for Adding Variables

This test is an alternative to Ramsey's RESET test for detecting specification errors in a model.

In order to explain this test procedure, here also we are using the examples of cost functions. Let $Y_i = \lambda_1 + \lambda_2 X_i + u_i$ [illegible] be a linear cost function and

$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$ [illegible] be a cubic cost function, where

Y = Total cost and X = output

If we now compare the linear cost function with the cubic cost function, then the linear cost function will be a restricted version of cubic cost function. The restricted regression $Y_i = \alpha_1 + \alpha_2 X_i + u_i$ assumes that the coefficients of the squared and cubed output terms, i.e., β_3 and β_4 are equal to zero.

To test this, the LM test proceeds as follows:

(i) We estimate the restricted regression $Y_i = \alpha_1 + \alpha_2 X_i + u_i$ by OLS method and obtain the residuals, $\hat{u}_i$.

(ii) If in fact the unrestricted regression $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$ is the true regression, the residuals obtained from restricted regression $Y_i = \alpha_1 + \alpha_2 X_i + u_i$ should be related to the squared and cubed output terms, i.e., X_i^2 and X_i^3.

(iii) This suggests that we regress the $\hat{u}_i$ obtained in step (i) on all the regressors (including those in the restricted regression), which in the present case means

$$\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^3 + v_i$$

where v_i is an error term with usual properties.

(iv) For large sample size, [illegible] shown that n (the sample size) times R^2 estimated from the regression [illegible] (5) follows a chi-square distribution with degrees of freedom equal to the number of restrictions imposed by the restricted regression, 2 in the present example since the terms X_i^2 and X_i^3 are dropped from the model.

Symbolically we write $nR^2 \sim \chi^2_{(2)}$ (4) (no. of restrictions = 2)

[illegible]

[illegible]

where [illegible] = total [illegible]

[illegible]

Although [illegible]

[illegible] for this purpose.

EXERCISE

1. What [illegible] mean by specification problems. What is meant by diagnostic checking of a model? Explain with the help of an example.
2. What do you mean by model selection? What are the [illegible] of selecting a good model?
3. What do you mean by specification errors? Give a brief account of different types of specification errors.
4. What do you mean by [illegible] specification errors? What are the consequences of these specification errors?
5. What do you mean by model specification errors? Explain the consequences of model specification errors when a relevant variable is omitted from a regression model.
6. What do you mean by model [illegible] errors? Explain the consequences of model specification errors when an irrelevant variable is included in a regression model.
7. What do you mean by errors of measurement of variables in a regression model? How can we detect the presence of measurement errors in a regression model?
8. What do you mean by [illegible] variable from a regression model? What are the different tests used for detecting the omitted variables [illegible] a model or [illegible] model.

[illegible]

On the assumption that [illegible]

[illegible]

(ii) [illegible] function?

Suppose the true model is [illegible]

[illegible]

(a) Find [illegible] and [illegible]

Show that [illegible]

Suppose the true model is [illegible] but we fit the model [illegible] and thus reveal the specification error of including an unnecessary variable in the model.

Show that the OLS estimators of the parameters of the incorrect model are all unbiased and consistent.

[illegible]

(ii) β_2 is statistically significant [illegible] but β_3 is not

[illegible]

The model [illegible] was estimated [illegible] observations. The results were

[illegible]

The same model was estimated with the restriction $\beta_1 = \beta_2$. Estimates were

[illegible]

(a) Suppose [illegible] were dropped from the equation. [illegible]

(b) Should the [illegible] be dropped?

21. The following table shows the values of expenditure on clothing (Y), total expenditure (X_1) and the price of clothing (X_2)

	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
X_2	16	13	10	9	7	5	4	3	3.5	2
X_1	15	20	30	42	50	54	65	72	85	90
Y	3.5	4.3	5	6	7	9	8	10	12	14

(i) Estimate the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

(ii) Estimate the model : $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$

(iii) If $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ is the true model, then examine the consequences on the regression parameters when X_2 is omitted from the model.

22. The following results were obtained from a sample of size 12

$\Sigma Y_i = 753$, $\Sigma Y_i^2 = 48{,}139$, $\Sigma X_{1i} Y_i = 40830$

$\Sigma X_{1i} = 643$, $\Sigma X_{1i}^2 = 34843$, $\Sigma X_{2i} Y_i = 6{,}796$

$\Sigma X_{2i} = 106$, $\Sigma X_{2i}^2 = 976$, $\Sigma X_{1i} X_{2i} = 5{,}779$

(i) Estimate the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

(ii) Estimate the model : $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$

(iii) Examine the impact on the regression parameters when X_2 is omitted from the true model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

APPENDIX

STATISTICAL TABLES

TABLE I

ORDINATES AND AREAS OF THE DISTRIBUTION OF STANDARD NORMAL VARIABLE*

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	61	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	62	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	63	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	5948349	65	[illegible]	7421539	[illegible]	[illegible]	[illegible]
25	[illegible]	5987063	66	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
26	[illegible]	[illegible]	67	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
27	3846627	6064199	68	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28	[illegible]	6102612	69	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
29	3825146	6140919	70	3122539	[illegible]	111	[illegible]	[illegible]
30	3813878	6179114	71	[illegible]	[illegible]	112	[illegible]	[illegible]
31	3802264	6217195	72	[illegible]	[illegible]	113	[illegible]	[illegible]
32	3790305	6255158	73	[illegible]	[illegible]	114	[illegible]	[illegible]
33	3778007	6293000	74	[illegible]	[illegible]	115	[illegible]	[illegible]
34	3765372	6330717	75	[illegible]	[illegible]	116	[illegible]	[illegible]
35	3752403	6368307	76	[illegible]	[illegible]	117	[illegible]	[illegible]
36	3739106	6405764	77	[illegible]	[illegible]	118	[illegible]	[illegible]
37	3725483	6443088	78	[illegible]	[illegible]	119	[illegible]	[illegible]
38	3711549	6480273	79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
39	3697277	[illegible]	80	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABLE I (Contd.)

x	φ(x)	Φ(x)	x	φ(x)	Φ(x)	x	φ(x)	Φ(x)
1.22	[illegible]	[illegible]	1.71	[illegible]	[illegible]	2.21	[illegible]	[illegible]
1.23	[illegible]	[illegible]	1.72	[illegible]	[illegible]	2.22	[illegible]	[illegible]
1.24	[illegible]	[illegible]	1.73	[illegible]	[illegible]	2.23	[illegible]	[illegible]
1.25	[illegible]	[illegible]	1.74	[illegible]	[illegible]	2.24	[illegible]	[illegible]
1.26	[illegible]	[illegible]	1.75	[illegible]	[illegible]	2.25	[illegible]	[illegible]
1.27	[illegible]	[illegible]	1.76	[illegible]	[illegible]	2.26	[illegible]	[illegible]
1.28	[illegible]	[illegible]	1.77	[illegible]	[illegible]	2.27	[illegible]	[illegible]
1.29	[illegible]	[illegible]	1.78	[illegible]	[illegible]	2.28	[illegible]	[illegible]
1.30	[illegible]	[illegible]	1.79	[illegible]	[illegible]	2.29	[illegible]	[illegible]
1.31	[illegible]	[illegible]	1.80	[illegible]	[illegible]	2.30	[illegible]	[illegible]
1.32	[illegible]	[illegible]	1.81	[illegible]	[illegible]	2.31	[illegible]	[illegible]
1.33	[illegible]	[illegible]	1.82	[illegible]	[illegible]	2.32	[illegible]	[illegible]
1.34	[illegible]	[illegible]	1.83	[illegible]	[illegible]	2.33	[illegible]	[illegible]
1.35	[illegible]	[illegible]	1.84	[illegible]	[illegible]	2.34	[illegible]	[illegible]
1.36	[illegible]	[illegible]	1.85	[illegible]	[illegible]	2.35	[illegible]	[illegible]
1.37	[illegible]	[illegible]	1.86	[illegible]	[illegible]	2.36	[illegible]	[illegible]
1.38	[illegible]	[illegible]	1.87	[illegible]	[illegible]	2.37	[illegible]	[illegible]
1.39	[illegible]	[illegible]	1.88	[illegible]	[illegible]	2.38	[illegible]	[illegible]
1.40	[illegible]	[illegible]	1.89	[illegible]	[illegible]	2.39	[illegible]	[illegible]
1.41	[illegible]	[illegible]	1.90	[illegible]	[illegible]	2.40	[illegible]	[illegible]
1.42	[illegible]	[illegible]	1.91	[illegible]	[illegible]	2.41	[illegible]	[illegible]
1.43	[illegible]	[illegible]	1.92	[illegible]	[illegible]	2.42	[illegible]	[illegible]
1.44	[illegible]	[illegible]	1.93	[illegible]	[illegible]	2.43	[illegible]	[illegible]
1.45	[illegible]	[illegible]	1.94	[illegible]	[illegible]	2.44	[illegible]	[illegible]
1.46	[illegible]	[illegible]	1.95	[illegible]	[illegible]	2.45	[illegible]	[illegible]
1.47	[illegible]	[illegible]	1.96	[illegible]	[illegible]	2.46	[illegible]	[illegible]
1.48	[illegible]	[illegible]	1.97	[illegible]	[illegible]	2.47	[illegible]	[illegible]
1.49	[illegible]	[illegible]	1.98	[illegible]	[illegible]	2.48	[illegible]	[illegible]
1.50	[illegible]	[illegible]	1.99	[illegible]	[illegible]	2.49	[illegible]	[illegible]
1.51	[illegible]	[illegible]	2.00	[illegible]	[illegible]	2.50	[illegible]	[illegible]
1.52	[illegible]	[illegible]	2.01	[illegible]	[illegible]	2.51	[illegible]	[illegible]
1.53	[illegible]	[illegible]	2.02	[illegible]	[illegible]	2.52	[illegible]	[illegible]
1.54	[illegible]	[illegible]	2.03	[illegible]	[illegible]	2.53	[illegible]	[illegible]
1.55	[illegible]	[illegible]	2.04	[illegible]	[illegible]	2.54	[illegible]	[illegible]
1.56	[illegible]	[illegible]	2.05	[illegible]	[illegible]	2.55	[illegible]	[illegible]
1.57	[illegible]	[illegible]	2.06	[illegible]	[illegible]	2.56	[illegible]	[illegible]
1.58	[illegible]	[illegible]	2.07	[illegible]	[illegible]	2.57	[illegible]	[illegible]
1.59	.1127042	[illegible]	2.08	[illegible]	[illegible]	2.58	[illegible]	[illegible]
1.60	[illegible]	[illegible]	2.09	[illegible]	[illegible]	2.59	[illegible]	[illegible]
1.61	[illegible]	[illegible]	2.10	[illegible]	[illegible]	2.60	[illegible]	[illegible]
1.62	[illegible]	[illegible]	2.11	[illegible]	[illegible]	2.61	[illegible]	[illegible]
1.63	[illegible]	[illegible]	2.12	[illegible]	[illegible]	2.62	[illegible]	[illegible]
1.64	[illegible]	[illegible]	2.13	[illegible]	[illegible]	2.63	[illegible]	[illegible]
1.65	[illegible]	[illegible]	2.14	[illegible]	[illegible]	2.64	[illegible]	[illegible]
1.66	[illegible]	[illegible]	2.15	[illegible]	[illegible]	2.65	[illegible]	[illegible]
1.67	[illegible]	[illegible]	2.16	[illegible]	[illegible]	2.66	[illegible]	[illegible]
1.68	[illegible]	[illegible]	2.17	[illegible]	[illegible]	2.67	[illegible]	[illegible]
1.69	[illegible]	[illegible]	2.18	.0370589	[illegible]	2.68	[illegible]	[illegible]
1.70	[illegible]	[illegible]	2.19	.0362615	.9857379	2.69	[illegible]	[illegible]

TABLE I (Contd.)

x	φ(x)	Φ(x)	x	φ(x)	Φ(x)	x	φ(x)	Φ(x)
2.72	[illegible]	[illegible]	3.01	[illegible]	[illegible]	3.30	[illegible]	[illegible]
2.73	[illegible]	[illegible]	3.02	[illegible]	[illegible]	3.31	[illegible]	[illegible]
2.74	[illegible]	[illegible]	3.03	[illegible]	[illegible]	3.32	[illegible]	[illegible]
2.75	[illegible]	[illegible]	3.04	[illegible]	[illegible]	3.33	[illegible]	[illegible]
2.76	[illegible]	[illegible]	3.05	[illegible]	[illegible]	3.34	[illegible]	[illegible]
2.77	[illegible]	[illegible]	3.06	[illegible]	[illegible]	3.35	[illegible]	[illegible]
2.78	[illegible]	[illegible]	3.07	[illegible]	[illegible]	3.36	[illegible]	[illegible]
2.79	[illegible]	[illegible]	3.08	[illegible]	[illegible]	3.37	[illegible]	[illegible]
2.80	[illegible]	[illegible]	3.09	[illegible]	[illegible]	3.38	[illegible]	[illegible]
2.81	[illegible]	[illegible]	3.10	[illegible]	[illegible]	3.39	[illegible]	[illegible]
2.82	[illegible]	[illegible]	3.11	[illegible]	[illegible]	3.40	[illegible]	[illegible]
2.83	[illegible]	[illegible]	3.12	[illegible]	[illegible]	3.41	[illegible]	[illegible]
2.84	[illegible]	[illegible]	3.13	[illegible]	[illegible]	3.42	[illegible]	[illegible]
2.85	[illegible]	[illegible]	3.14	[illegible]	[illegible]	3.43	[illegible]	[illegible]
2.86	[illegible]	[illegible]	3.15	[illegible]	[illegible]	3.44	[illegible]	[illegible]
2.87	[illegible]	[illegible]	3.16	[illegible]	[illegible]	3.45	[illegible]	[illegible]
2.88	[illegible]	[illegible]	3.17	[illegible]	[illegible]	3.46	[illegible]	[illegible]
2.89	[illegible]	[illegible]	3.18	[illegible]	[illegible]	3.47	[illegible]	[illegible]
2.90	[illegible]	[illegible]	3.19	[illegible]	[illegible]	3.48	[illegible]	[illegible]
2.91	[illegible]	[illegible]	3.20	[illegible]	[illegible]	3.49	[illegible]	[illegible]
2.92	[illegible]	[illegible]	3.21	[illegible]	[illegible]	3.50	[illegible]	[illegible]
2.93	[illegible]	[illegible]	3.22	[illegible]	[illegible]	3.51	[illegible]	[illegible]
2.94	[illegible]	[illegible]	3.23	[illegible]	[illegible]	3.52	[illegible]	[illegible]
2.95	[illegible]	[illegible]	3.24	[illegible]	[illegible]	3.53	[illegible]	[illegible]
2.96	[illegible]	[illegible]	3.25	[illegible]	[illegible]	3.54	[illegible]	[illegible]
2.97	[illegible]	[illegible]	3.26	[illegible]	[illegible]	3.55	[illegible]	[illegible]
2.98	[illegible]	[illegible]	3.27	[illegible]	[illegible]	3.56	[illegible]	[illegible]
2.99	[illegible]	[illegible]	3.28	[illegible]	[illegible]	3.57	[illegible]	[illegible]
3.00	[illegible]	[illegible]	3.29	[illegible]	[illegible]	3.58	[illegible]	[illegible]

* Abridged from Table 1 of *Biometrika Tables for Statisticians*, vol. 1, with the kind permission of the Biometrika Trustees.

TABLE II

DISTRIBUTION OF STANDARD NORMAL VARIABLE

Values of τ_α

α	0.05	0.025	0.01	0.005
τ_α	1.645	1.960	2.326	2.576

TABLE III
χ^2 DISTRIBUTION*
VALUES OF χ^2_{α}

α / n	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

For larger values of n, the quantity $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a standard normal variable.

*Abridged from Table 8 of *Biometrika Tables for Statisticians*, vol. 1, with the kind permission of the Biometrika Trustees.

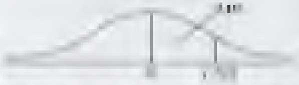
TABLE IV
t-DISTRIBUTION*
VALUES OF $t_{\alpha,\nu}$

Example

Pr(t > 2.086) = 0.025

Pr(t > 1.725) = 0.05 for df = 20

Pr(|t| > 1.725) = 0.10



Pr / df	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note: The smaller probability shown at the head of each column is the area in one tail; the larger probability is the area in both tails.

Source: From E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 12, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.

TABLE V

F-DISTRIBUTION*

TABLE V (Contd.)

TABLE VI
THE DURBIN-WATSON d-STATISTIC
SIGNIFICANCE POINTS OF d_L AND d_U : 5%

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.61	1.40	—	—	—	—	—	—	—	—
7	0.70	1.35	0.46	1.89	—	—	—	—	—	—
8	0.76	1.33	0.55	1.77	0.36	2.28	—	—	—	—
9	0.82	1.32	0.62	1.69	0.45	2.12	0.29	2.58	—	—
10	0.87	1.32	0.69	1.64	0.52	2.01	0.37	2.41	0.24	2.82
11	0.92	1.32	0.65	1.60	0.59	1.92	0.44	2.28	0.31	2.64
12	0.97	1.33	0.81	1.57	0.65	1.86	0.51	2.17	0.37	2.50
13	1.01	1.34	0.86	1.56	0.71	1.81	0.57	2.09	0.44	2.39
14	1.04	1.35	0.90	1.55	0.76	1.77	0.63	2.03	0.50	2.29
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
150	1.72	1.74	1.70	1.76	1.69	1.77	1.67	1.78	1.66	1.80
200	1.75	1.77	1.74	1.78	1.73	1.79	1.72	1.81	1.71	1.82

Note : k' = Number of explanatory variables excluding the constant.